Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy

Brian D. Ziebart

December 2010
CMU-ML-10-110

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee
J. Andrew Bagnell, Co-chair
Anind K. Dey, Co-chair
Martial Hebert
Dieter Fox, University of Washington

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

This research was sponsored by an R. K. Mellon Fellowship and by the National Science Foundation under contract no. EEEC-0540865. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

Keywords: Machine learning, decision making, probabilistic modeling, maximum entropy, inverse optimal control, influence diagrams, informational revelation, feedback, causality, goal inference

Abstract

Predicting human behavior from a small amount of training examples is a challenging machine learning problem. In this thesis, we introduce the principle of maximum causal entropy, a general technique for applying information theory to decision-theoretic, game-theoretic, and control settings where relevant information is sequentially revealed over time. This approach guarantees decision-theoretic performance by matching purposeful measures of behavior (Abbeel & Ng, 2004), and/or enforces game-theoretic rationality constraints (Aumann, 1974), while otherwise being as uncertain as possible, which minimizes worst-case predictive log-loss (Grünwald & Dawid, 2003).

We derive probabilistic models for decision, control, and multi-player game settings using this approach. We then develop corresponding algorithms for efficient inference that include relaxations of the Bellman equation (Bellman, 1957), and simple learning algorithms based on convex optimization. We apply the models and algorithms to a number of behavior prediction tasks. Specifically, we present empirical evaluations of the approach in the domains of vehicle route preference modeling using over 100,000 miles of collected taxi driving data, pedestrian motion modeling from weeks of indoor movement data, and robust prediction of game play in stochastic multi-player games.

For Emily

Acknowledgments

I owe much gratitude to many people for their support and encouragement leading to this thesis. First and foremost, I thank my thesis advisors, Anind Dey and Drew Bagnell, for directing my research towards exciting problems and applications, and for supplying me with enough guidance and knowledge to stay on the right path, but also enough freedom to feel empowered by the journey. I also thank Martial Hebert and Dieter Fox for serving on my thesis committee and providing their valuable insights on my research.

Many thanks go to the Quality of Life Technology Center (National Science Foundation Grant No. EEEC-0540865) and the R.K. Mellon Foundation for supporting this research and the broader goal of creating assistive technologies for those who would benefit from them.

I am indebted to Roy Campbell and Dan Roth for advising my initial experiences with computer science research at the University of Illinois, as well as Manuel Roman and the rest of the Systems Research Group, for taking me in and making my experience enjoyable. That work continues to motivate my thinking.

Without Eric Oatneal and Jerry Campolongo of Yellow Cab Pittsburgh, and Maury Fey of Westinghouse SURE, our data collections and studies of driver preferences would not have been possible.

I thank my co-authors with whom I had the pleasure of working on the Purposeful People Prediction project at Carnegie Mellon/Intel: Nathan Ratliff, Garratt Gallagher, Christoph Mertz, Sidd Srinivasa, and Kevin Peterson. Making robots reason and function purposefully is definitely a team activity.

Being a member of the Machine Learning Department (and the wider SCS and CMU community) has been a wonderful experience. I thank the faculty, the students, and the staff for making it a collegial and intellectually stimulating environment for learning and exploring this exciting field. Special thanks go to Andrew Arnold, Anind's research group, Andy Carlson, Lucia Castellanos, Hao Cen, Miro Dudík, Joey Gonzalez, Sue Ann Hong, Jonathan Huang, Andreas Krause, LairLab, Thomas LaToza, Jure Leskovec, Mary McGlohon, Abe Othman, and Diane Stidle.

I thank my officemates over the years—Hao Cen, Robert Fisher, Hai-Son Le, Jonathan Moody, and Ankur Parikh—who were patient when maximizing entropy extended beyond my research to my desk.

I thank Andrew Maas for his late night coding, ideas, karaoke singing, and friendship in our endeavors together.

My former housemates Jure Leskovec, Andrew Arnold, and Thomas LaToza, made the transition to Pittsburgh and Carnegie Mellon a pleasant one filled with cookies and a more than occasional games of Settlers of Catan.

To my parents: nature or nurture, I owe both to you. I thank you and my sister, Sarah, for fueling my curiosity, valuing my education, and for making me the person who I am today.

To my wife, Emily, your patience and support through the ups and downs of the research cycle have been instrumental.

Contents

List of Figures

List of Tables

List of Algorithms

Part I

Preliminaries

Chapter 1

Introduction

"The future influences the present just as much as the past."
— Friedrich Nietzsche (Philosopher, 1844–1900).

As humans, we are able to reason about the future consequences of our actions and their relationships to our goals and objectives—even in the presence of uncertainty. This ability shapes most of our high-level behavior. Our actions are typically *purposeful* and sensitive to the *revelation of new information* in the future; we are able to anticipate the possible outcomes of our potential actions and intelligently select appropriate actions that lead to desirable results. In fact, some psychologists have defined intelligence itself as "goal-directed adaptive behavior" (Sternberg & Salter, 1982). This reasoning is needed not only as a basis for intelligently choosing our own behaviors, but also for being able to infer the intentions of others, their probable reactions to our own behaviors, and rational possibilities for group behavior.

We posit that to realize the long-standing objective of artificial intelligence—the creation of computational automata capable of reasoning with "human-like" intelligence—, those automata will need to possess similar goal-directed, adaptive reasoning capabilities. This reasoning is necessary for enabling a robot to intelligently choose its behaviors, and, even more importantly, to allow the robot to infer and understand, by observation, the underlying reasons guiding intelligent behavior in humans. The focus of our work is on constructing predictive models of goal-directed adaptive behaviors that enable computational inference of a person's future behavior and long-term intentions. Importantly, prediction techniques must be robust to differences in context and should support the transfer of learned behavior knowledge across similar settings. We argue that just as goal-directedness and adaptive reasoning are critical components of intelligent behaviors, they must similarly be central components of our predictive models for those behaviors.

For **prescriptive** models (*i.e.*, those that provide optimal decisions), rich planning and decision-making frameworks that incorporate both goals and adaptive reasoning exist. For example, in a Markov decision process (Puterman, 1994), both goal-directedness and adaptive reasoning are incorporated by an optimal controller, which selects actions that maximize the expected utility over future random outcomes. Though these models are useful for control purposes where the provided

optimal action can simply be executed, they are often not useful for predictive purposes because observed behavior is rarely absolutely and consistently optimal[1]. Similarly, game-theoretic solution concepts specify joint rationality requirements on multi-player behavior but lack the uniqueness to be able to predict what strategy players will employ.

Instead, **predictive** models capable of forecasting future behavior by estimating the probabilities of future actions are needed. Unfortunately, many existing probabilistic models of behavior have very little connection to planning and decision-making frameworks. They instead consider behavior as a sequence of random variables without considering the context in which the behavior is situated, and how it relates to the available options for efficiently satisfying the objectives of the behavior. Thus, the existing approaches lack the crucial goal-directedness and adaptive reasoning that is characteristic of high-level human behavior. These existing models can still be employed to predict goal-directed adaptive behavior despite being neither inherently goal-directed nor incorporating adaptive reasoning themselves. However, the mismatch with the properties of high-level behavior comes at a cost: slower rates of learning, poorer predictive accuracy, and worse generalization to novel behaviors and decision settings.

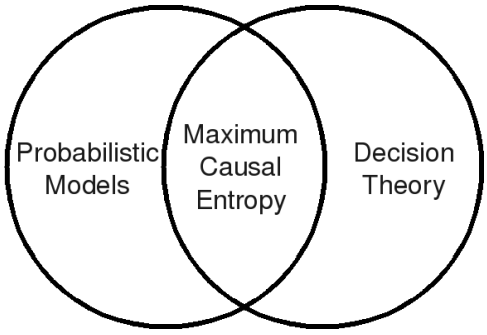

Figure 1.1: A Venn diagram representing probabilistic models, decision-theoretic models and their intersection, where our maximum causal entropy approach for forecasting behavior resides.

In this thesis, we develop predictive models of goal-directed adaptive behavior by explicitly incorporating goals and adaptive reasoning into our formulations. We introduce the **principle of maximum causal entropy**, our extension to the maximum entropy principle that addresses settings where side information (from *e.g.*, nature, random processes, or other external uncertain influences) is sequentially revealed. We apply this principle to existing decision-theoretic and strategic reasoning frameworks to obtain predictive probabilistic models of behavior that possess goal-directed and adaptive properties. From one perspective, this work generalizes existing probabilistic graphical model techniques to the sequentially revealed information settings common in decision-theoretic and game-theoretic settings. From a second perspective, this work generalizing existing optimal

[1] A notable exception is the duality of control and estimation in the linear quadratic setting established by Kalman (1960).

control techniques in decision-theoretic and strategic frameworks from prescribing optimal actions to making predictions about behavior with predictive guarantees. This high-level combination of probabilistic graphical models and decision theory is depicted in Figure 1.1.

We now make the central claim of this thesis explicit:

> **The principle of maximum causal entropy creates probabilistic models of decision making that are purposeful, adaptive, and/or rational, providing more accurate prediction of human behavior.**

To validate this claim, we introduce the principle of maximum causal entropy, employ it to derive probabilistic models that are inherently purposeful and adaptive, develop efficient algorithms for inference and learning to make those models computationally tractable, and apply those models and algorithms to behavior modeling tasks.

1.1 Contributions to the Theory of Behavior Prediction

The main contributions of this thesis to the support of the theory of behavior prediction in support of the central thesis statement are as follows:

- **The principle of maximum causal entropy** (Ziebart et al., 2010b) extends the maximum entropy framework (Jaynes, 1957) to settings with information revelation and feedback, providing a general approach for modeling observed behavior that is purposeful, adaptive, and/or rational.

- **Maximum causal entropy inverse optimal control** (Ziebart et al., 2010b) resolves ambiguities in the problem of recovering an agent's reward function from demonstrated behavior (Ng & Russell, 2000; Abbeel & Ng, 2004), and generalizes conditional random fields (Lafferty et al., 2001) to settings with dynamically-revealed side information.

- **Maximum causal entropy inverse linear-quadratic regulation** (Ziebart et al., 2010b) resolves the special case of recovering the quadratic utility function that best explains sequences of continuous controls in linear dynamics settings.

- **Maximum entropy inverse optimal control** (Ziebart et al., 2008b) is the special case of the maximum causal entropy approach applied to settings with deterministic state transition dynamics.

- **Maximum causal entropy influence diagrams** (Ziebart et al., 2010b) expand the maximum causal entropy approach to settings with addition uncertainty over side information, such as learning to model diagnostic decision making. This approach resolves the question of recovering reward for influence diagrams that explain demonstrated behavior.

- **Maximum causal entropy correlated equilibria** Ziebart et al. (2010a) extend **maximum entropy correlated equilibria** (Ortiz et al., 2007) for normal-form games (*i.e.*, single-shot), which provide predictive guarantees for jointly rational multi-player settings, to the dynamic, sequential game setting.

1.2 Motivation

If technological trends hold, we will see a growing number of increasingly powerful computational resources available in our everyday lives. Embedded computers will provide richer access to streams of information and robots will be afforded a greater level of control over our environments, while personal and pervasive networked devices will always make this information and control available at our fingertips.

In our view, whether these technologies become a consistent source of distraction or a natural extension of our own abilities (Weiser, 1991) depends largely on their algorithmic ability to understand our behavior, predict our future actions and infer our intentions and goals. Only then will our computational devices and systems be best able to augment our own natural capabilities. The possible benefits of technologies for behavior prediction are numerous and include:

- Accurate predictive models of the ways we interact with and control our surrounding computational resources can be employed to automate those interactions on our behalf, reducing our interaction burden.

- A knowledge of our current intentions and goals can be used to filter irrelevant, distracting information out, so that our computational systems only provide information that is pertinent to our current activities and intentions.

- Systems that can understand our intentions can help guide us in achieving them if we begin to err or are uncertain, essentially compensating for our imperfect control or lack of detailed information that our computational systems may possess.

While systems with improved abilities to reason about human behavior have applicability across a wide range of domains for the whole spectrum of users, we are particularly motivated by the problem of assisting older people in their daily lives. We envision a wealth of assistive technologies that augment human abilities and assist users in living longer, more independent lives.

1.3 Purposefulness, Adaptation, and Rationality

Part of the central claim of this thesis is that by designing probabilistic models that inherently possess the same high-level properties as intelligent behavior, we can achieve more accurate predictions of that behavior. Identifying and defining all the properties that characterize intelligent

behavior is an important task that has previously been investigated. We now review the previously identified characteristics of behavior and use them to motivate the perspective of this thesis.

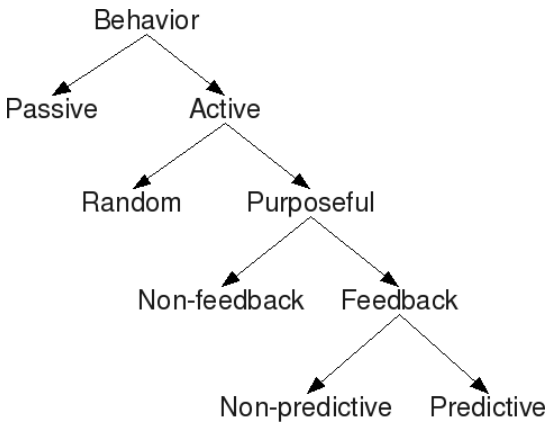

Figure 1.2: Rosenblueth et al. (1943)'s hierarchy of types of behavior. Categories at greater depth in this hierarchy are considered to be related to reasoning capabilities associated with greater intelligence.

Rosenblueth et al. (1943) provide a hierarchy of different classes of behavior (Figure 1.2). They distinguish between **active** and **passive** behavior based on whether the actor or agent is a source of energy. Active behavior is either directed towards some target (**purposeful**) or undirected (**random**). Our first formal definition adopts Taylor (1950)'s broader conceptualization of purposefulness, which incorporates both means and ends.

Definition 1.1 (Taylor, 1950)**.** ***Purposefulness*** *is defined as follows:*

> *There must be, on the part of the behaving entity, i.e., the agent: (a) a desire, whether actually felt or not, for some object, event, or state of affairs as yet future; (b) the belief, whether tacit or explicit, that a given behavioral sequence will be efficacious as a means to the realization of that object, event, or state of affairs; and (c) the behavior pattern in question. Less precisely, this means that to say of a given behavior pattern that it is purposeful, is to say that the entity exhibiting that behavior desires some goal and is behaving in a manner it believes appropriate to the attainment of it. (Taylor, 1950)*

This definition captures many of the characteristics that one might associate with intelligent behavior (Sternberg & Salter, 1982). Namely, that it is based on an ability to reason about the effects that the combination of current and future actions have towards achieving a long-term goal or set of objectives. Actions that are counter-productive or very inefficient in realizing those long-term goals are avoided in favor of actions that efficiently lead to the accomplishment of intended objectives. Similarly, myopic actions that provide immediate gratification are avoided if they do

not provide future benefits across a longer horizon of time. Note that behavior sequences must be *efficacious* in realizing a goal rather than optimal. We argue the theoretic, algorithmic, and empirical benefits of incorporating purposefulness into predictive models of behavior throughout this thesis.

Rosenblueth et al. (1943) further divide purposeful behavior into three categories:

- **Non-feedback**, which is not influenced by any feedback;
- **Non-predictive feedback**, which responds to feedback when it is provided; and
- **Predictive feedback**, which incorporates beliefs about anticipated future feedback.

We provide broader definitions than those shaped by the dominant problem of focus for those authors (automatic weapon targeting).

Definition 1.2. ***Non-predictive feedback-based behavior*** *is characterized by the influence of feedback of any type received while the behavior is executed.*

This definition better matches our employed definition of purposefulness (Definition 1.1) by allowing for feedback relating both to the agent's goal and to changes in the "appropriateness" of manners for attaining it.

Definition 1.3. ***Predictive feedback-based behavior*** *is characterized by the influence of anticipated feedback to be received in the future on current behavior.*

Consider the example of a cat pursuing a mouse to differentiate predictive and non-predictive feedback-based behavior (Rosenblueth et al., 1943). Rather than moving directly towards the mouse's present location (non-predictive), the cat will move to a position based on its belief of where the mouse will move (predictive). In this thesis we will narrowly define *adaptive behavior* to be synonymous with predictive feedback-based behavior. We will restrict our consideration to the anticipatory setting in this thesis, but we note that the simpler non-anticipatory setting can be viewed as a sequence of non-feedback-based behaviors with varying inputs.

We illustrate the differences relating to adaptation by using the following example shown in Figure 1.3. There are three routes that lead to one's home (Point C). The two shortest routes share a common path up until a point (B). A trusted source knows that there has been an accident on one of the two shortest routes. Any driver that attempts to take that particular route will be ensnared in traffic delays lasting for hours. Depending on the information provided and the problem setting, the side information about traffic congestion falls into each of the three settings:

1. **Complete availability.** The trusted source reveals exactly which bridge is congested. One should then choose to take the other (non-congested) short route home.

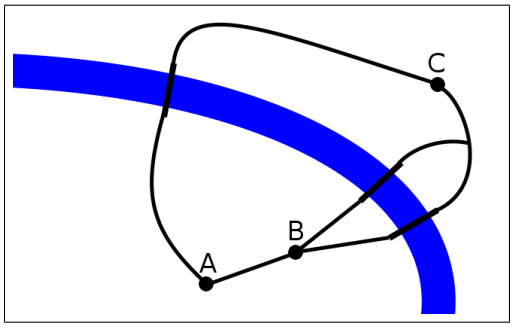

Figure 1.3: An example decision problem with three routes crossing a river to connect a starting location (A) to a destination (C). One of the two rightmost routes, which are components of the fastest routes, is known to be congested. Choosing a route depends crucially on whether the congested route can be detected at a shared vantage point (B) before having to commit to the possibly congested point.

2. **Complete latency.** The trusted source can only reveal that the traffic accident occurred on one of the two shorter routes. Based on the road network topology, it is impossible to determine which bridge is congested before committing to a route. Choosing the third route that is known to be delay-free is likely to be preferable in this setting.

3. **Information revelation.** The trusted source again can only reveal that the traffic accident occurred on one of the two shorter routes. However, there is a vantage point at B that allows one to observe both bridges. In this case, taking the shared route to point B and then deciding which route to take based on available observations is likely to be preferable.

The key distinction between these settings is when the relevant information about the congested bridge becomes available, and specifically whether it is known before all decisions are made (Setting 1), after all decisions are made (Setting 2), or in-between decisions (Setting 3). The last setting, which we refer to as **information revelation**, provides the opportunity for adaptive behavior primary concern in this thesis. High-level behavior and our models of it should not only respond to revealed information, but also anticipate what might be revealed when choosing preceding actions. In addition to being simply revealed over time, *what* side information is revealed can be a function of the behavior executed up to that point, and *the value* of that side information can potentially be influenced by that behavior as well.

The sophistication of the observed behavior goes far deeper than the first-order and second-order dynamics models originally considered by Rosenblueth et al. (1943). Adversarial situations exist where the acting agent is aware of being observed and acts in accordance to this knowledge using his, her, or its own model of the observer's capabilities and intentions. When the awareness of observer and observee's knowledge and intentions is common, behavior takes the form of a game, and in the limit of infinitely recursive rationality, game-theoretic equilibrium solution concepts can be employed. These concepts provide criteria for assessing the rationality of players' strategies.

1.4 Prescriptive Versus Predictive Models

Frameworks for planning and decision making provide the necessary formalisms for representing purposeful, adaptive behavior. However, these frameworks are generally employed for **prescriptive** applications where, given the costs of states and actions (or some parameters characterizing those costs), the optimal future cost-minimizing action for every situation is computed. Importantly, for **predictive** applications, any decision framework should be viewed as an approximation to the true motives and mechanics of observed behavior. With this in mind, observed behavior is typically not consistently optimal for any fixed prescriptive model parameters for many reasons:

- Discrepancies between the features employed by the *observer* modeling behavior, and the *observed* generating behavior may exist.
- Observed behavior may be subject to varying amounts of control error that make generating perfectly optimal behavior impossible.
- Additional factors that are either unobservable or difficult to accurately model may influence observed behavior.

As a result, the "best" actions according to an optimal controller often do not perfectly predict actual behavior. Instead, probabilistic models that allow for sub-optimal behavior and uncertainty in the costs of states and actions are needed to appropriately predict purposeful behavior. We create such probabilistic, predictive models from prescriptive planning and decision frameworks in this thesis.

1.5 Maximum Causal Entropy

The principle of maximum entropy (Jaynes, 1957) is a powerful tool for constructing probability distributions that match known properties of observed phenomena, while not committing those distributions to any additional properties not implied by existing knowledge. This property is assured by maximizing Shannon's **information entropy** of the probability distribution subject to constraints on the distribution corresponding to existing knowledge.

In settings with information revelation, future revealed information should not *causally influence* behavior occurring earlier in time. Doing so would imply a knowledge of the future that violates the temporal revelation of information imposed by the problem setting. Instead, the distribution over the revealed information rather than the particular instantiation of the revealed information can and should influence behavior occurring earlier in time. For example, in hindsight a person's decision to carry an umbrella may seem strange if it did not rain during the day, but given a forecast with a 60% chance of rain when the decision was made, the decision would be reasonable.

Based on the distinction between belief of future information variables and their actual instantiation values, we present **the principle of maximum causal entropy** as an extension of the general principle of maximum entropy to the information revelation setting. It enables the principle of maximum entropy to be applicable in problems with partial observability, feedback, and stochastic influences from nature (*i.e.*, stochastic dynamics), as well as imperfect recall (*e.g.*, information upon which past decisions were based is forgotten) and game theoretic settings.

1.6 Applications and Empirical Evaluations

We validate our approach by evaluating it on a number of sequential decision prediction tasks using models we develop in this thesis based on the principle of maximum causal entropy:

- **Inverse optimal control** models that recover the reward or utility function that explains observed behavior.
- **Maximum causal entropy influence diagrams** that extend the approach to settings with additional imperfect information about the current state of the world.
- **Maximum causal entropy correlated equilibria** that extend the approach of the thesis to sequential, multi-player settings where deviation regrets constrain behavior to be jointly rational.

We apply these models to a number of prediction tasks.

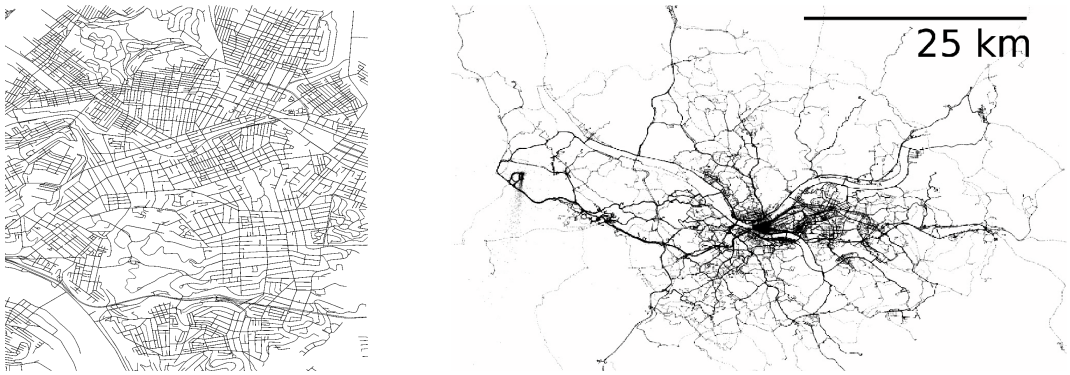

Figure 1.4: A portion of the Pittsburgh road network representing the routing decision space (left) and collected positioning data from Yellow Cab Pittsburgh taxi drivers (right).

In our first application, we learn the preferences of drivers navigating through a road network from GPS data (Figure 1.4). We employ the learned model for personalized route recommendations and for future route predictions using Bayesian inference methods. These resulting predictions enable systems to provide relevant information to drivers.

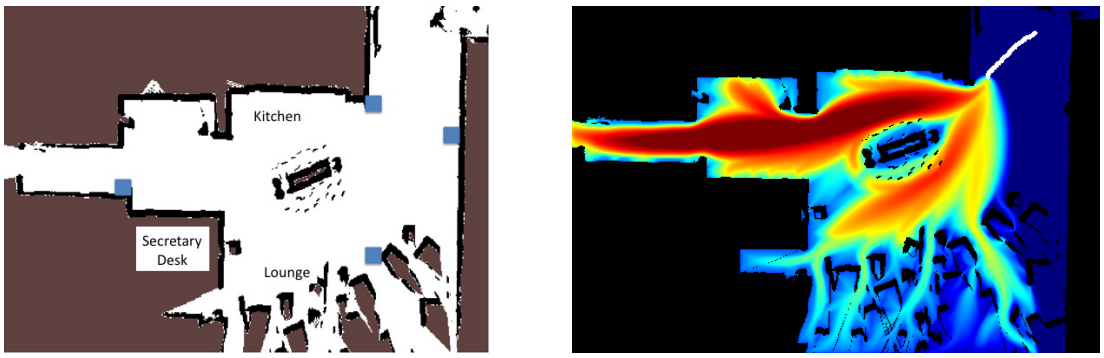

Figure 1.5: A portion of the Intel Pittsburgh laboratory (left) and future trajectory predictions within that environment (right).

In our second application, we learn to predict the trajectories of moving people within an environment from LADAR data (Figure 1.5). These prediction enable robots to generate more complementary motion trajectories that reduce the amount of hindrance to people.

Finally, we present a set of smaller experiments to demonstrate the range of applications for the approach. We apply maximum causal entropy correlated equilibria for multi-agent robust strategy prediction. We employ maximum causal entropy influence diagrams to predict actions in a partially observed, inverse diagnostics application. Lastly, we demonstrate inverse linear quadratic regulation for helicopter control.

1.7 Thesis Organization and Reader's Guide

This thesis is organized into five parts: Preliminaries, Theory, Algorithms, Applications, and Conclusions. Descriptions of each part and chapter of the thesis are as follows:

Part I, Preliminaries: Motivations and related work for the behavior forecasting task and review of the background material of techniques that are employed in the thesis	
Chapter 1	Motivations for the behavior prediction task and discussion of its purposeful and adaptive characteristics, which are leveraged in this thesis to make useful predictions of behavior based on small amounts of training data
Chapter 2	Review of the decision making frameworks that pose behavior as a utility-maximizing interaction with a stochastic process; Review of probabilistic graphical models
Chapter 3	Review of inference and learning for decision making from the perspectives of probabilistic graphical models, optimal control theory, and discrete choice theory; Discussion of limitations of those approaches

Part II, Theory: The principle of maximum causal entropy for decision-theoretic and game-theoretic settings	
Chapter 4	Review of information theory for quantifying uncertainty, and the introduction of causal information theory—the extension of information theory to settings with interaction and feedback—and its existing results and applications
Chapter 5	Review of the principle of maximum entropy as a tool for constructing probability distributions that provide predictive guarantees; Introduction of the principle of maximum causal entropy, which extends those predictive guarantees to the interactive setting
Chapter 6	Application of the maximum causal entropy principle to the problem of inverse optimal control where a "softened" Markov decision problem's reward function is learned that best explains observed behavior
Chapter 7	Introduction of the maximum entropy influence diagram, a general-purpose framework for approximating conditional distributions with sequentially revealed side information
Chapter 8	Extension of the maximum causal entropy approach to multi-player, non-cooperative, sequential game settings with the jointly rational maximum causal entropy correlated equilibria

Part III, Algorithms: Probabilistic inference, convexity-based learning, and Bayesian latent variable inference	
Chapter 9	Efficient algorithms for inference in maximum causal entropy models based primarily on a "softened" interpretation of the Bellman equations and analogs to efficient planning techniques
Chapter 10	Gradient-based learning algorithms for maximum causal entropy models that leverage the convexity properties of the optimization formulation
Chapter 11	Efficient algorithms for inferring goals and other latent variables within maximum causal entropy models from partial behavior traces

Part IV, Applications: Behavior learning and prediction tasks	
Chapter 12	Learning route selection decisions of drivers in road networks to support prediction and personalization tasks; Comparisons to inverse optimal control and directed graphical model approaches
Chapter 13	Forecasting motions of pedestrians to more intelligently plan complementary robot routes that are sensitive to hindering those pedestrians; Comparisons to particle-based simulation approaches
Chapter 14	Multi-player strategy prediction for Markov games; Modeling of behavior in partially observable settings for inverse diagnostics application; Predicting continuous control for helicopter hovering

Part V, Conclusions: Open questions and concluding thoughts	
Chapter 15	A set of possible future extensions incorporating models of perception, learning pay-offs from demonstrated multi-player strategies, and interactive prediction-based applications
Chapter 16	A concluding summary of the thesis and some final thoughts

Additionally, proofs of the theorems from throughout the thesis are presented in Appendix A.

A sequential reading of the thesis provides detailed motivation, exploration of background concepts, theory development from general to specific formulations, and, lastly, algorithms and empirical justifications in the form of applications. However, the central contributions of the thesis can be understood by the following main points:

- Frameworks for decision making processes, such as the Markov decision process (Section 2.2.1), view behavior as the interaction of an agent with a stochastic process. Purposefulness and adaption are incorporated into "solutions" to problems in these frameworks that prescribe the optimal action to take in each state (*i.e.*, a policy) that maximizes an expected reward based on possible random future states.

- Existing approaches to the behavior forecasting task either learn the policy rather than the much more generalizable reward function (Section 3.1), or do not appropriately incorporate uncertainty when learning the reward (Section 3.2.1).

- The principle of maximum entropy (Section 5.1) is a general approach for learning that has a number of intuitive justifications and it provides important predictive guarantees (e.g., Theorem 5.2) based on information theory (Section 4.1). From it, modern graphical models (*e.g.*, Markov random fields and conditional random fields (Section 2.1.2)) for estimating probability distributions are obtained (Section 5.1.3).

- Extension of the maximum entropy approach using causal information theory (Section 4.2) in the form of the principle of maximum causal entropy (Section 5.2) is required for the maximum entropy approach to be applicable to the sequential interaction setting.

The remainder of the thesis develops the principle of maximum causal entropy approach for a number of settings, formulates algorithms for inference and learning in those settings, and applies the developed approach on prediction tasks. Portions of the thesis may be of greater interest based on the background knowledge of reader. We highlight some of interesting themes to possibly follow:

- For those already familiar with inverse optimal control, Section 3.2 reviews existing inverse optimal control approaches and criticisms of those approaches. Section 6.2 establishes the properties of the maximum causal entropy that address those criticisms. A comparison of inverse optimal control techniques on a navigation prediction task is presented in Section 12.6.2.

- For machine learning theorists desiring to understand the relationship of maximum causal entropy to conditional random fields (CRFs): Section 4.1 and Section 5.1 up to 5.1.3 provide a derivation of CRFs using the principle of maximum entropy. Section 5.2 then provides the generalization to information revelation settings, and Chapter 6 focuses on the modeling setting that is the natural parallel to chain CRFs where side information is sequentially revealed.

- For those with an optimal control background: Section 6.2.2 provides a key interpretation of maximum causal entropy inference (Chapter 9) as a softened version of the Bellman equation for a parametric Markov decision process (Definition 2.10 in Section 2.2.1). The learning problem (Theorem 6.4 and Chapter 10) is that of finding the parameters that best explain observed behavior.

- For game theorists, Chapter 8 employs maximum causal entropy to provide unique correlated equilibria for Markov games with strong predictive guarantees. Chapter 5 provides the general formulation of the principle of maximum causal entropy this approach is based upon. Section 14.1 studies the predictive benefits of these equilibria on randomly generated games.

Portions of this thesis have previously appeared as workshop publications: Ziebart et al. (2007), Ziebart et al. (2009a), Ziebart et al. (2010a); and as conference publications: Ziebart et al. (2008b), Ziebart et al. (2008c), Ziebart et al. (2008a), Ratliff et al. (2009), Ziebart et al. (2009b) Ziebart et al. (2010b).

Chapter 2

Background and Notation

"No sensible decision can be made any longer without taking into account not only the world as it is, but the world as it will be."

— Isaac Asimov, (Writer, 1920–1992).

Behavior sequence forecasting with the principle of maximum causal entropy relies upon many existing concepts and ideas from information theory and artificial intelligence. Specifically, the major theoretical contribution of this thesis is the extension of the principle of maximum entropy to sequential information settings commonly represented using decision-theoretic frameworks. This extension also generalizes existing graphical models. We review existing probabilistic graphical models, which are also commonly employed for prediction tasks, and the decision-theoretic models of planning and decision making that we rely upon in this thesis. Lastly, we introduce some of the notations and summarize the terminology that we employ throughout this thesis.

2.1 Probabilistic Graphical Models

We now review existing techniques for modeling random variables using probabilistic graphical models. We develop comparison approaches and baseline evaluations for our applications using these techniques throughout this thesis.

2.1.1 Bayesian Networks

Bayesian networks are a framework for representing the probabilistic relationships between random variables.

Definition 2.1. *A **Bayesian network**, $BN = (G, P)$, is defined by:*

- *A directed acyclic graph, G, that expresses the structural (conditional) independence relationships between variables, $X \in \mathcal{X}$; and*

- *Conditional probability distributions that form the joint probability distribution,* $\mathrm{P}(\boldsymbol{X})$*, based on the parent variables of each variable in the graph, parents*(X)*, as follows:*

$$\mathrm{P}(\boldsymbol{X}) = \prod_i \mathrm{P}(X_i | \textit{parents}(X_i)). \tag{2.1}$$

A great deal of research has been conducted on efficient methods for inferring the probability distribution of a set of variables in the Bayesian network given a set of observed variables.

One of the most attractive properties of Bayesian networks is the simplicity of learning model parameters. Given fully observed data, the maximum likelihood conditional probability distributions are obtained by simply counting the joint occurrences of different combinations of variables. Overfitting to a small amount of observed data can be avoided by adding a Dirichlet (*i.e.*, *pseudo-count*) prior to these counts so that unobserved combinations will have non-zero probability.

Bayesian networks are typically extended to temporal settings as dynamic Bayesian networks (Definition 2.2) by assuming that the structure and conditional probabilities are stationary over time.

Definition 2.2. *A* ***dynamic Bayesian network*** *(Murphy, 2002),* $DBN = (G, P)$ *is defined by:*

- *A template directed graphical structure, G; and*
- *A set of conditional probability distributions defining the relationship for* $P(\boldsymbol{X}_{t+1}|\boldsymbol{X}_t)$ *(where* $\boldsymbol{X}_t = \{X_{t,1}, ..., X_{t,N}\}$ *and* $\boldsymbol{X}_{t+1} = \{X_{t+1,1}, ..., X_{t+1,N}\}$*).*

The joint probability for a time sequence of variables is obtained by repeating this structure over a fixed time horizon, T*:* $\boldsymbol{X}_{1:T} = \prod_{t=1}^{T} P(\boldsymbol{X}_{t+1}|\boldsymbol{X}_t)$*.*

Conceptually, a Bayesian network is obtained by "unrolling" the DBN to a specific time-step size. We will leverage this same concept of abstractly representing a model over a few timesteps.

2.1.2 Markov and Conditional Random Fields

Markov random fields (MRFs) (Kindermann et al., 1980) represent the *synergy* between combinations of variable values rather than their conditional probability distributions.

Definition 2.3. *A* ***Markov random field*** *for variables* $\boldsymbol{X}$*,* $MRF = (G, F, \theta)$*, is defined by:*

- *An undirected graph, G, that specifies cliques,* C_j*, over variables;*
- *Feature functions (*$F = \{f \rightarrow \mathbb{R}^K\}$*) and model parameters (*$\theta \in \mathbb{R}^K$*) that specify potential functions,* $\theta_{C_j}^{\top} \; f_{C_j}(X_{C_j})$*, over variable cliques.*

The corresponding probability distribution is of the form:

$$\mathrm{P}(\boldsymbol{X}) \propto e^{\sum_j \theta_{C_j}^\top f_{C_j}(\boldsymbol{X}_{C_j})},$$

where C_j are all cliques of variables in G.

Parameters of a MRF, $\{\theta_{C_j}\}$, are generally obtained through convex optimization to maximize the probability of observed data, rather than being obtained from a closed-form solution.

The conditional generalization of the MRF is the conditional random field (CRF) (Lafferty et al., 2001). It estimates the probability of a set of variables conditioned on another set of **side information variables**.

Definition 2.4. *A **conditional random field**, $CRF = (G, F, \theta)$ is defined by:*

- *An undirected graph, G, of cliques between conditioned variables, **Y**, and side information variables, **X**;*
- *Feature functions ($f \rightarrow \mathbb{R}^K$) and model parameters ($\theta \in \mathbb{R}^K$) that specify potential functions, $\theta_{C_j}^\top \ f_{C_j}(Y_{C_j}, X_{C_j})$, over variable cliques.*

The conditional probability distribution with feature functions over cliques of X and Y values is of the form:

$$\mathrm{P}(\boldsymbol{Y}|\boldsymbol{X}) \propto e^{\sum_j \left(\theta_{C_j}^\top f_j(X_{C_j}, Y_{C_j})\right)}. \tag{2.2}$$

Generally, any graph structure can be employed for a CRF. However, chain models (Definition 2.5) are commonly employed for temporal or sequence data.

Definition 2.5. *A **chain conditional random field** is a conditional random field over time-indexed variables where the cliques are:*

- *Over consecutive label variables, $C_j(Y_t, Y_{t+1})$; or*
- *Over intra-timestep variables, $C_j(X_t, Y_t)$.*

The conditional probability distribution is of the form:

$$\mathrm{P}(\boldsymbol{Y}|\boldsymbol{X}) \propto e^{\sum_t \sum_k (\theta_k f_k(X_t, Y_t) + \phi_k g_k(Y_t, Y_{t+1}))}. \tag{2.3}$$

In a number of recognition tasks, the additional variables of a conditional random field are observational data, and the CRF is employed to recognize underlying structured properties from these observations. This approach has been successfully applied to recognition problems for text (Lafferty et al., 2001), vision (Kumar & Hebert, 2006), and activities (Liao et al., 2007a; Vail et al., 2007).

Conditional random fields generalize Markov random fields to conditional probability settings where a set of side information data is available. The contribution of this thesis can be viewed as the generalization of conditional random fields to settings where the side information variables are not immediately available. Instead, those variables are assumed to be revealed dynamically over time.

2.2 Decision-Theoretic Models

A number of frameworks for representing decision-making situations have been developed with the goal of appropriately representing the factors that influence a decision and then enabling the optimal decision to be efficiently ascertained. Here we review a few common ones. Importantly, all of these frameworks pose behavior as a sequence of interactions with a stochastic process that maximize expected utility. We leverage these *prescriptive* decision-theoretic frameworks to create *predictive* decision models later in this thesis. The principle of maximum causal entropy is what enables the appropriate application of probabilistic estimation techniques to the sequential interaction setting.

2.2.1 Markov Decision Processes

One common model for discrete planning and decision making is the Markov decision process (MDP), which represents a decision process in terms of a graph structure of **states** and **actions** (Figure 2.1a and Figure 2.1c), rewards associated with those graph elements, and stochastic transitions between states.

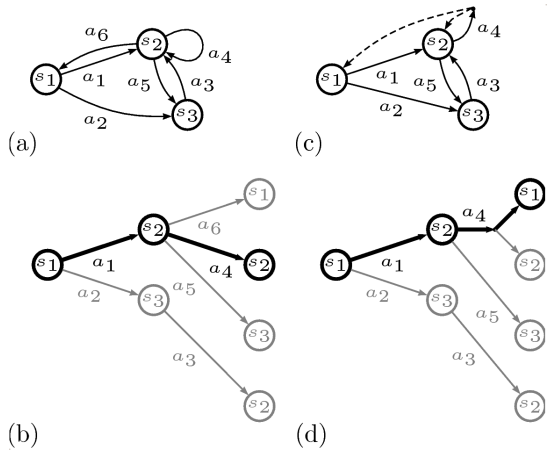

Figure 2.1: (a) The transition dynamics of a Markov decision process with only deterministic state transitions. (b) The tree of possible states and actions after executing two actions (starting in s_1) for this deterministic MDP. (c) The transition dynamics of a Markov decision process with stochastic state transitions. (d) The tree of possible states and actions after executing two actions (starting in s_1) for this stochastic MDP.

Definition 2.6. *A **Markov decision process** (MDP) is a tuple, $\mathcal{M}_{MDP} = (S, A, \mathrm{P}(s'|s, a), R(s, a))$, of:*

- *A set of **states** ($s \in S$);*

- *A set of **actions** ($a \in A$) associated with states;*
- *Action-dependent **state transition dynamics** probability distributions ($\mathrm{P}(s'|s,a)$) specifying a next state (s'); and*
- *A **reward function** ($R(s,a) \to \mathbb{R}$).*

At each timestep t, the state (S_t) is generated from the transition probability distribution (based on S_{t-1} and A_{t-1}) and observed before the next action (A_t) is selected.

A **trajectory** through the MDP consists of sequences of states and actions such as those shown in bold in Figure 2.1b and 2.1d. We denote the trajectory as $\zeta = \{\mathbf{s}_{1:T}, \mathbf{a}_{1:T}\}$. It has an associated cumulative reward $R(\zeta) = \sum_{t:s_t,a_t \in \zeta} \gamma^t R(s_t, a_t)$. The optional **discount factor**, $1 \geq \gamma > 0$, makes the reward contribution of future states and actions to the cumulative reward less significant than the current one. It can be interpreted as modifying the transition dynamics of the MDP to have a $1-\gamma$ probability of terminating after each time step. The remaining transition probabilities are scaled by a complementary factor of γ.

Optimal policies

The MDP is "solved" by finding a deterministic **policy** ($\pi(s) \to A$) specifying the action for each state that yields the highest expected cumulative reward, $\mathrm{E}_{\mathrm{P}(\mathbf{s}_{1:T},\mathbf{a}_{1:T})}[\sum_{t=0}^{T} \gamma^t R(s_t, a_t)|\pi]$ over a finite time horizon, T, or an infinite time horizon (Bellman, 1957).

Theorem 2.7. *The optimal action policy can be obtained by solving the Bellman equation,*

$$\pi(s) = \operatorname*{argmax}_a \left\{ R(s,a) + \gamma \sum_{s'} \mathrm{P}(s'|s,a) V(s') \right\} \tag{2.4}$$

$$V^*(s) = \max_a \left\{ R(s,a) + \gamma \sum_{s'} \mathrm{P}(s'|s,\pi(s)) V^*(s') \right\}. \tag{2.5}$$

Alternately, the optimal **state value function**, $V^*(s)$ can be defined in terms of the optimal **action value function**, $Q^*(s,a)$:

$$V^*(s) = \max_a \{R(s,a) + Q^*(s,a)\}$$
$$Q^*(s,a) = \gamma \sum_{s'} \mathrm{P}(s'|s,a) V^*(s').$$

This definition will be useful for understanding the differences of the maximum causal entropy approach and its algorithms for obtaining a stochastic policy.

The Bellman equations can be recursively solved by updating the $V^*(s)$ values (and policies, $\pi(s)$) iteratively using dynamic programming. The **value iteration** algorithm (Bellman, 1957)

iteratively updates $V^*(s)$ by expanding its definition to be in terms of $V^*(s')$ terms, $V^*(s) = \max_a R(s,a) + \gamma \sum_{s'} \mathrm{P}(s'|s,a)V^*(s')$, rather than separately obtaining a policy. The **policy iteration** algorithm applies Equation 2.4 to obtain a policy and then repeatedly applies the updates of Equation 2.5 until convergence, and then repeats these two steps until no change in policy occurs. We refer the reader to Puterman (1994)'s overview of MDPs for a broader understanding of their properties and relevant algorithms.

Stochastic and mixed policies

Optimal control methods tend to focus on deterministic policies, $\pi(s)$, that are **Markovian** (*i.e.*, conditionally independent of past states and actions given the current state). This is sensible since there always exists a deterministic policy that maximizes the expected reward. However, more general classes of policies exist that are of interest in this thesis. In a **stochastic policy**, each action is chosen according to a probability distribution, $\pi(a|s) \in [0,1]$, rather than deterministically. Both **stationary** (time-independent) and **non-stationary** (time-dependent) policies exist that depend on the state (and timestep if non-stationary) but no additional variables. More generally, a class of **mixed** policies require additional memory. For example, in a mixture of deterministic policies, a distribution over different deterministic policies, $\{\pi(s)_i\}$, is weighted by a set of probabilities, $\{\lambda_i\}$. One of these policies is sampled and then actions from it are (deterministically) executed to create a trajectory of actions and states.

Theorem 2.8 (Special case of Feinberg & Shwartz (2002), Theorem 6.1)**.** *Let $\pi^{(1)}, \pi^{(2)}, ...$ be an arbitrary sequence of policies and $\lambda_1, \lambda_2, ...$ be a sequence of scalars such that $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$. The randomized Markov policy π defined by*

$$\pi_t(a|s) \triangleq \frac{\sum_{i=1}^{\infty} \lambda_i \mathrm{P}(A_t = a, S_t = s|\pi^{(i)})}{\sum_{i=1}^{\infty} \lambda_i \mathrm{P}(S_t = s|\pi^{(i)})}, t \geq 0 \tag{2.6}$$

whenever the denominator in Equation 2.6 is not equal to 0. Then, for all $t \geq 0$, s, and a,

$$\mathrm{P}(A_t = a, S_t = s|\pi) = \sum_{i=1}^{\infty} \lambda_i \mathrm{P}(A_t = a, S_t = s|\pi^{(i)}). \tag{2.7}$$

As a consequence of Theorem 2.8, any mixed policy has a stochastic policy that, in expectation, has the same number of state-action occurrences. If the policies being mixed, $\pi^{(1)}, \pi^{(2)}, ...$, are stationary, the resulting stochastic policy, π, will be as well. This result is often used to argue that, at least in the sense of expected state-action executions, mixed policies afford no greater generality than stochastic policies (Feinberg & Shwartz, 2002).

Partial observability

An important extension of the MDP is to settings with uncertain states. In many domains, the full state, S, of the decision problem may only partially be known. **Partially observable Markov decision processes** (POMDP) allow for planning in this uncertain setting (Drake, 1962).

Definition 2.9. *A **partially observable Markov decision process** (POMDP) is a tuple, $\mathcal{M}_{POMDP} = (S, A, O, \mathrm{P}(s'|s,a), \mathrm{P}(o|s), R(s,a))$, of **states** ($s \in S$), **actions** ($a \in A$), **observations** ($o \in O$), action-dependent **state transition dynamics** probability distributions ($\mathrm{P}(s'|s,a)$) specifying a next state (s'), state-dependent **observation dynamics** ($\mathrm{P}(o|s)$) and a **reward function** ($R(s,a) \to \mathbb{R}$). At each timestep t, the state (S_t) is generated from the transition probability distribution (based on S_{t-1} and A_{t-1}), but only the observation variable, O_t, distributed according to the state is observed before the next action (A_t) is selected.*

The POMDP (Definition 2.9) can be viewed as a MDP that is augmented with observation variables and state-dependent observation variable dynamics. Instead of the state being revealed to the agent before each action is selected, as in the MDP, only the observation variable is revealed. Thus, the agent must operate with only partial knowledge of its current state.

Parametric reward functions

In this thesis, we are often be concerned with the setting where the rewards of the Markov decision process are linearly parameterized according to some parameters, θ. Though we do not explicitly assume this formulation, it follows as a consequence of the principle of maximum causal entropy. Specifically, we assume that each state-action pair (or state depending on the type of reward) has an associated vector of features and the reward that it provides is a linear function of those features as shown in Equation 2.8.

$$R_\theta(s,a) = \theta^\top \mathbf{f}_{s,a} \tag{2.8}$$

Definition 2.10. *A **parametric-reward Markov decision process** (PRMDP) is defined as a tuple:*

$$\mathcal{M}_{PRMDP} = (S, A, \mathrm{P}(s'|s,a), \{\boldsymbol{f}_{s,a}\}, \theta). \tag{2.9}$$

The rewards of states and actions in this setting are defined according to Equation 2.8 based on parameter vector θ.

When all of the parameters of a PRMDP (Definition 2.10) are known, it can be readily interpreted as a standard Markov decision process (Definition 2.6). The main advantage of this formulation is experienced when the parameter vector is much smaller than the sets of states and actions. In that case, the parameter vector transfers very easily to other PRMDPs that have the same space of features. Much of the focus of this thesis is the setting where the reward parameters are unknown. This setting of an PRMDP without reward function is denoted as $\mathcal{M}_{\text{PRMDP}}/\theta$.

2.2.2 Linear-Quadratic Control

Tractable extension of optimal control frameworks to continuous state and continuous action settings is often difficult. This is because integration over continuous variables is required and those integrations rarely admit closed-form solutions. A notable and important class of exceptions for problems of control is the linear-quadratic setting of Definition 2.11.

Definition 2.11. *A **linear quadratic regulation** setting is characterized by state dynamics that are distributed according to a linear function of past state and action variables with Gaussian noise. Cost functions are quadratic and linear in the state variables*[1] *parameterized by* $\boldsymbol{Q}$ *and* $\boldsymbol{R}$ *respectively, as shown in Equation 2.10:*

$$\begin{aligned} &\boldsymbol{s}_{t+1} = A\boldsymbol{s}_t + B\boldsymbol{a}_t + \epsilon_t \\ &\epsilon_t \sim N(0, \Sigma) \\ &\textit{Cost}(\boldsymbol{s}_{1:T}, \boldsymbol{a}_{1:T}) = \sum_{t=1}^{T} \boldsymbol{s}_t^\top Q \boldsymbol{s}_t + \boldsymbol{s}_t^\top R, \end{aligned} \tag{2.10}$$

where $\boldsymbol{Q}$ *is a symmetric, positive semi-definite matrix. More formally, the linear-quadratic control setting is described as a tuple of values,* $\mathcal{M}_{LQ} = \{A, B, \Sigma, Q, R\}$*, containing all of the previously described transition dynamic and cost parameters.*

The optimal state-dependent control for this setting, $\pi^*(\mathbf{s}) \rightarrow a \in \mathcal{A}$, can be solved in closed form. A similar linear-quadratic control formulation exists for continuous-time settings (also with closed-form solution), but we shall only consider the discrete time setting in this thesis. We will also consider the setting with an LQ model with unknown cost parameters, denoted $\mathcal{M}_{\text{LQ}}/\{Q, R\}$.

2.2.3 Influence Diagrams

Influence diagrams (IDs) (Miller et al., 1976; Howard & Matheson, 1984) are a graphical representation of uncertainty. They can be viewed as a generalization of the more familiar Bayesian networks that includes decision making and value assessments in addition to the uncertainty nodes of Bayes nets. Structurally, an influence diagram is a directed acyclic graph with three types of nodes:

- **Uncertainty nodes** that represent random variables with probability distributions that are conditioned on their parent variables.
- **Decision nodes** that represent decision outcomes that are made with the knowledge of the values of the node's parent variables.
- **Value nodes** that represent additive utility functions that are dependent on the node's parent variables.

Together the interaction of these variables can be used to represent decision processes with varying amounts of information availability. Note that a Bayesian network is simply the special case of an influence diagram consisting only of uncertainty nodes.

[1] Quadratic and linear costs for the action variables are also possible, but, without loss of generality, the past action can be added to the state and any action costs represented as state-based costs.

Definition 2.12. *Formally, an influence diagram is specified by a tuple* $\mathcal{M}_{ID} = \{U, D, V, E, P_U : par(U), U \rightarrow [0,1], f_V : par(V) \rightarrow \mathbb{R}\}$ *of the three different types of nodes (U, D, V), directed edges (E), conditional probability distributions for all uncertainty nodes (*P_U*), and value functions for all value nodes (*f_V*).*

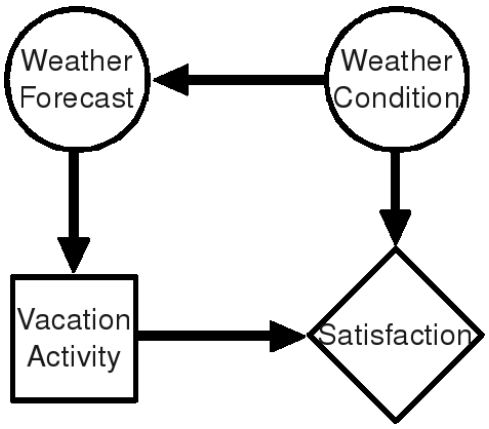

Figure 2.2: An illustrative influence diagram representation (Shachter, 2007) of making vacation decisions based on weather, which can only be indirectly observed through a noisy weather forecast.

A canonical influence diagram is shown in Figure 2.2 to illustrate the relationships of variables within a vacation decision task. In this decision problem, the future weather conditions are known only indirectly to the decision maker through a weather forecast. The decision maker must select a vacation activity with a satisfaction utility that depends on the actual weather conditions and the vacation activity selected.

Influence diagram optimal policies

The general decision task in an influence diagram is to select conditional value assignments for the decision nodes (*i.e.*, deterministic policies) so that the expected combined value of all value nodes is maximized:

$$\pi_D^*(\text{par}(D)) \triangleq \underset{\pi_D(\text{par}(D))}{\text{argmax}} \, \text{E}_{D,U}\left[\sum_V f_V(\text{par}(V))|\pi(\text{par}(D))\right]. \tag{2.11}$$

A number of methods for obtaining this optimal policy exist. Influence diagrams were originally solved by unrolling the graphical structure into decision trees and computing expectations of value nodes over the uncertainty nodes, while maximizing decision nodes upward from the leafs of the tree (Howard & Matheson, 1984). Shachter (1986) provides a variable elimination technique that iteratively removes nodes by marginalizing over uncertainty nodes and collapsing individual decision nodes.

Relationship to Markov decision processes

Influence diagrams are quite powerful at representing a whole range of decision-making tasks. They can be used to represent Markov decision processes (Shachter & Peot, 1992), as shown in Figure 2.3.

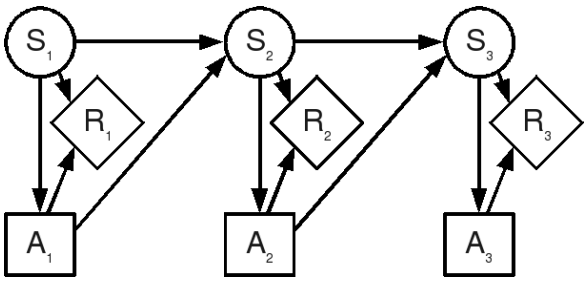

Figure 2.3: An influence diagram representation of a Markov decision process.

At each timestep t, there is an uncertainty node for the state (S_t) that is observed, a decision node that represents the action (A_t) to be selected, and a value node for the MDP's reward (R_t) that is received. The edges of the influence diagram indicate the influences on each node. Each next state is determined by the previous state and action's values. The reward is based on the state and action (or more simply just the state in some MDPs). Finally, the action is determined based only on the current state.

Influence diagrams can similarly be employed to represent partially-observable Markov decision processes, but only by using a significantly more complex structure to represent the influences on each action, as shown in Figure 2.4.

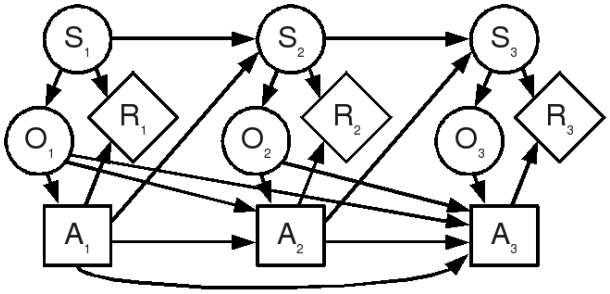

Figure 2.4: An influence diagram representation of a partially-observable Markov decision process.

An uncertainty node for an observation at each timestep is added to the model. The state is now unobserved and must be inferred based on these observations and the previous actions. This is denoted by each action's parents being all of the available observations and previous actions rather than the current state.

2.3 Notation and Terminology

We now introduce some of the notations employed frequently throughout this thesis. We usually employ the most compact notation possible, often suppressing information that can be understood from context (*e.g.*, the dimensions of vectors, the domains of probability distributions).

Variables, values, sets, and vectors of each:

- Random variables: X, Y, Z or $X_1, X_2, ..., X_n$
- Variable values: x or $x_1, x_2, ..., x_n$
- Sets of variables: $\{X, Y, Z\}$ or $\{X_i\}_{i=1:n} = \{X_i\}_{1:n} = \{X_i\} = \{X_1, X_2, ..., X_n\}$
- Sets of values: $\{x_i\}_{i=1:n} = \{x_i\}_{1:n} = \{x_i\} = \{x_1, x_2, ..., x_n\}$
- Vectors of variables: $\mathbf{X}_{1:n} = \mathbf{X} = (X_1\ X_2\ ...\ X_n)^\top$
- Vectors of values: $\mathbf{x}_{1:n} = \mathbf{x} = (x_1\ x_2\ ...\ x_n)^\top$
- Double-indexed vectors (matrices): $\mathbf{X}^{1:t}_{1:n} = (\mathbf{X}^1_{1:n}\ \mathbf{X}^2_{1:n}\ ...\ \mathbf{X}^t_{1:n})$

Probability distributions, expectations, and entropies:

- Probability distribution: $\mathrm{P}(X = x) = \mathrm{P}_X(x) = \mathrm{P}(x)$
- Conditional probability distribution: $\mathrm{P}(Y = y|X = x) = \mathrm{P}_{Y|X}(y|x) = \mathrm{P}(y|x)$
- Causally conditioned probability distribution: $P(\mathbf{Y}||\mathbf{X}) = \prod_i P(Y_i|\mathbf{X}_{1:i}, \mathbf{Y}_{1:i-1})$
- Expectation of function f: $\mathrm{E}_{P(x)}[f(x)] = \mathrm{E}[f(x)] = \sum_x \mathrm{P}(x)f(x)$ (or $\int_x \mathrm{P}(x)\ f(x)\ dx$)
- Entropy: $H_P(X) = H(X) = E[-\log P(x)] = -\sum_x P(x) \log P(x)$
- Causal entropy: $H_P(\mathbf{Y}||\mathbf{X}) = H(\mathbf{Y}||\mathbf{X}) = E[-\log P(\mathbf{Y}||\mathbf{X})]$

Note on "value" terminology

It is inevitable when drawing upon ideas from a number of different areas of research for conflicts in terminology to arise. We note one of these conflicts here in an attempt to avoid confusion.

In graphical models, a random variables takes on a value (*i.e.*, a particular instantiation). In optimal control problems, a state has a value (*i.e.*, an expected cumulative reward). Thus, when discussing random state variables, the value of the state variable can be ambiguous and confusing. We have attempted to refer to instantiations when we mean the former, and to value functions when we mean the latter.

Relative and causal entropy notation

Traditionally, the relative entropy is denoted as: $H(Y||X)$. However, this notation conflicts with the established causal entropy notation. Instead, we employ $H(Y//X)$ to denote the relative entropy.

2.4 Summary

In this chapter, we have provided an overview of the decision frameworks to which we will apply the developed principle of maximum causal entropy. Within these frameworks, behavior can be viewed as a utility-maximizing interaction with a stochastic process. This utility measure captures the purposefulness of behavior, while reasoning correctly about stochastic interaction is inherently adaptive. However, these frameworks (and algorithms for "solving them") are designed for prescribing behavior rather than predicting it.

Additionally, we have described existing directed and undirected probabilistic graphical models. Those graphical models have been previously employed for predicting sequences of behavior. A familiarity with them is therefore important for understanding the distinctions between the approach of this thesis and past approaches for predicting sequential data. In Chapter 5, we revisit the probabilistic graphical models from a maximum entropy perspective and from that perspective create a more general approach for settings with sequential information revelation. In Chapters 6 and 7 we apply this new approach to Markov decision processes and influence diagrams.

Chapter 3

Related Work

"One's first step in wisdom is to question everything -
and one's last step is to come to terms with everything."
— Georg Christoph Lichtenberg (Physicist, 1742–1799).

In this chapter, we review existing approaches for approximating and forecasting behavior developed under a variety of different names (*e.g.*, imitation learning, apprenticeship learning) and with a variety of specific techniques (*e.g.*, inverse optimal control, inverse reinforcement learning, robust KL control, Boltzmann action value distributions, and conditional logit models). These approaches have been shaped primarily by three general perspectives: probabilistic graphical models (Section 3.1), optimal control (Section 3.2), and discrete choice theory (Section 3.3) for three primary purposes: as a method of robust or efficient approximate inference, as a method for learning how to behave using demonstrated behavior (Argall et al., 2009), and as a tool for forecasting future behavior from past observed behavior. We discuss the similarities and differences between these existing approaches and our maximum causal entropy-based models at a high-level in Section 3.4 and in more detail throughout later chapters of this thesis.

3.1 Probabilistic Graphical Models for Decision Making

The development of inference and learning techniques for probabilistic graphical models has received a large amount of attention from the machine learning community over the past twenty years. Though influence diagrams (Section 2.2.3) generalize Bayesian networks (2.1.1), many of the advances in inference were developed specifically for Bayesian networks and not their more general decision-based counterpart. A steady line of research has attempted to apply efficient approximate inference techniques in non-decision-based probabilistic graphical models to influence diagrams, Markov decision processes, and other decision modeling frameworks (Cooper, 1988; Shachter & Peot, 1992; Jensen et al., 1994; Zhang, 1998; Attias, 2003; Toussaint & Storkey, 2006; Ardis & Brown, 2009; Kappen et al., 2010).

Recently, probabilistic graphical model approaches have been applied for learning, reasoning about, and predicting behavior. Most of these approaches have been concerned with the activity recognition problem of recognizing underlying activity sequences from noisy sensor data (Bao & Intille, 2004). However, some recent work has focused on the prediction of future decision making without the aid of noisy observations. The underlying probabilistic graphical model techniques employed are similar regardless of the availability of sensor data.

We review probabilistic graphical model approaches for both inference of optimal decision policies and learning to model demonstrated behavior to highlight the similarities and differences with our approach. We focus more heavily on models for learning as they better match our primary motivation in this thesis.

3.1.1 Directed Graphical Model Approaches

Policy learning for Markov decision processes

Directed graphical models have been employed to create probability distributions that represent behavior in Markov decision processes by directly estimating an observed policy, $\pi(a|s)$.

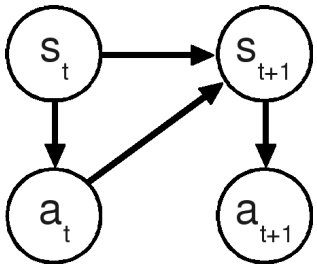

Figure 3.1: A simple two-slice dynamic Bayesian network model of decision making in a Markov decision process incorporating state (s) and action (a) variables.

A simple dynamic Bayesian network representation for a Markov decision process is shown in Figure 3.1. The distribution of actions for each state can be learned from empirical data, with the option of incorporating knowledge about these probabilities using a prior distribution. The state transition dynamics of the model, $\mathrm{P}(s_{t+1}|s_t, a_t)$, are typically assumed to be known from the Markov decision process. Often when using this model with different specified goal states ($s_g \in \mathcal{S}_g$), s_T (for some large T) is fixed to s_g, and inference over the remaining latent actions is performed to obtain the distribution of the next action.

In practice, the basic model is often augmented with additional variables (Verma & Rao, 2006) as shown in Figure 3.2. This augmented model enables behavior for differing goal states to be learned by incorporating a goal variable (g). Additional variables indicate whether the goal is reached (r) and represent observations (o) for partially-observable Markov decision processes.

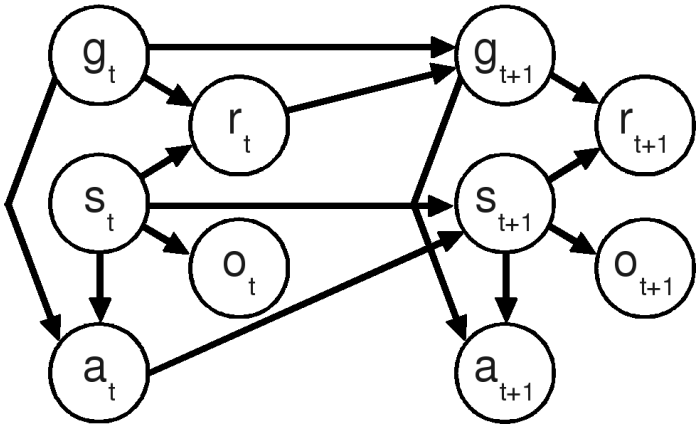

Figure 3.2: A more complex two-slice dynamic Bayesian network model of decision making in a Markov decision process. It incorporates variables for the goal (g), variables indicating whether the goal has been reached (r) and observation variables (o).

Verma & Rao (2006) employ this probabilistic model to infer the posterior probability of a latent goal from a partial sequence of actions and states.

Dynamic Bayesian networks are also often employed for activity recognition tasks (Bui et al., 2002; Tapia et al., 2004). Those models focus on the relationship between state variables and observations and often do not have explicit reward and activity variables. In theory, future observation variables can be marginalized over to provide predictions of future state. However, activity recognition techniques focus heavily on estimating the conditional relationships between observations and state, and not necessarily on accurately predicting future states without those observations. Thus, they should not be expected to predict future decision making data more accurately than those purposed for that task.

The main disadvantage of these dynamic Bayesian network approaches is their disconnect from the reward function of the MDP. The reward function provides a level of abstraction beyond the policy, explaining behavior not in terms of *what is happening*, but in terms of the underlying reasons *why it is happening*. For example, given the reward function, the optimal policy for any goal state can easily be obtained from the Bellman equation. By contrast, in this directed model, a goal-dependent policy must be learned for each possible goal. Therefore, the naïve directed approach does not easily provide the transfer of learned knowledge between the learned policies for different goals.

Toussaint & Storkey (2006) reduce this disconnect between dynamic Bayesian network models of decision making and reward-based Markov decision processes by formulating a reward-based model as a mixture of different length Markov chains of states and actions (Figure 3.3) with a reward variable conditioned on the final state and action of each chain: $\mathrm{P}(r_t|a_t, s_t) \triangleq \mathrm{reward}(a_t, s_t)$. They show that arbitrary MDPs can be embedded within this graphical model representation and

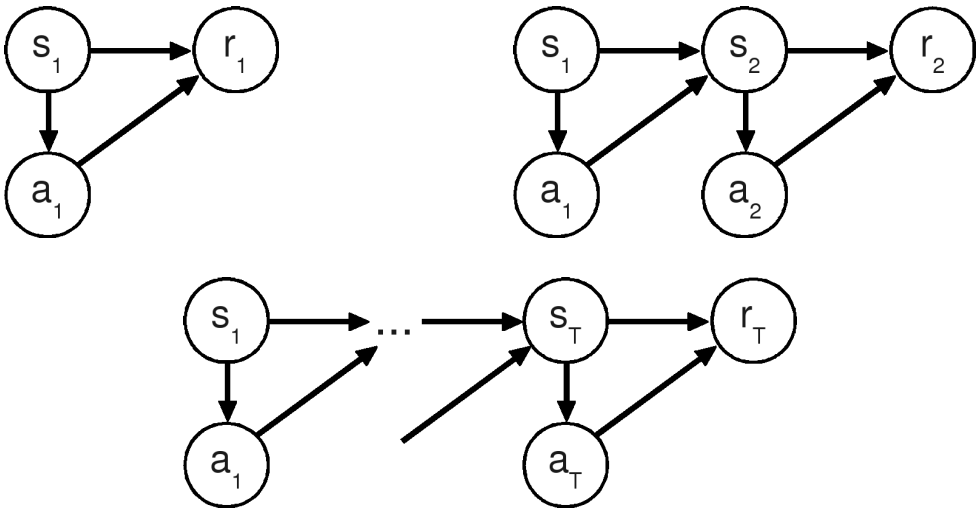

Figure 3.3: The Markov chains with terminal rewards that are mixed together to obtain a probabilistic graphical model inference procedure for a MDP's optimal policy.

that maximizing a conditional probability in that model equates to finding the maximum expected value policy of the corresponding (PO)MDP. This model, however, is only suitable for inference given known reward values, as no procedure is suggested for learning reward functions from demonstrated behavior.

Variable-length Markov models and memory-based approaches

The variable-length Markov model (VLMM) (Galata et al., 2001) is a higher-order dynamic Bayesian network built on the assumption that future intentions are captured by a longer-term history of states and actions. The distribution of next actions in the VLMM is conditioned on a sequence of previous actions where the length of the sequence depends on the availability of similar previously observed data:

$$\mathrm{P}(a_t|\mathbf{a}_{1:t-1}, \mathbf{s}_{1:t}) = \tilde{\mathrm{P}}\left(a_t|\mathbf{a}_{(t-k(\mathbf{a}_{1:t-1},\mathbf{s}_{1:t})):(t-1)}, \mathbf{s}_{(t-k(\mathbf{a}_{1:t-1},\mathbf{s}_{1:t}):t}\right), \tag{3.1}$$

where $k(\mathbf{a}_{1:t-1}, \mathbf{s}_{1:t-1}) \rightarrow \mathbb{Z} \geq 0$. More generally, memory-based and analogical approaches to planning (Veloso, 1994) stitch together sequences of previously observed behavior in an attempt to reasonably accomplish a new objective.

The main limitation of the VLMM and memory-based approaches is their dependence on a large amount of similar previously-observed behavior. In the absence of similar previous data, the VLMM degrades into a random walk with poor performance. Similarly, if previously employed plans with similar origin and goal states as a desired new plan are unavailable, the resulting "stitched-together" plan of a memory-based approach will tend to be biased away from direct plans to the goal and towards portions of the state spaces that have been frequently observed previously.

3.1.2 Undirected Graphical Approaches

Undirected graphical models (Section 2.1.2), and, specifically, the conditional random field (Lafferty et al., 2001), have experienced remarkable success in recognition tasks for text (Lafferty et al., 2001), machine vision (Kumar & Hebert, 2006), and activities. For activity recognition tasks, side information corresponds to observed characteristics (Liao et al., 2007a; Vail et al., 2007). Under this approach, a potential function is learned for consecutive states, Y_t and Y_{t+1}, but these sequential state potentials are only learned to the degree that they are needed to compensate for when state-observation prediction power alone is insufficient.

Ardis & Brown (2009) propose a conditional random field approach to decision inference where the optimal response $\mathbf{Y}^*_{\mathbf{X}}$ to any outcome of nature, $\mathbf{X} \sim \mathrm{P}(\mathbf{X})$, is obtained. Toussaint (2009) arrives at a model that requires a negative reward function, but is otherwise identical, using a directed graphical model approach where all behavior is conditioned on an additional set of variables, $\mathbf{z} = 1$, where $\mathrm{P}(z_t|y_t, x_t) = e^{-\mathrm{cost}(y_t,x_t)}$. Ziebart et al. (2008b) propose this same distribution as an approximation to a maximum entropy formulation of modeling decision making in a Markov decision process with stochastic dynamics. The joint distribution of responses is obtained using the MDP's state transition dynamics, and nature's outcomes are marginalized over to find the "optimal" conditional probability when $\mathbf{X}$ is latent. Expanding this model to the predictive setting, we have:

$$\mathrm{P}(Y_t|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t-1}) \propto \sum_{\mathbf{X}_{t+1:T}, \mathbf{Y}_{t+1:T}} e^{-\sum_t \mathrm{cost}(y_t,x_t)} \mathrm{P}(\mathbf{X}_{t+1:T}|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t-1}). \tag{3.2}$$

For some simple problems with stochastic state-transition dynamics, this approach yields a policy with optimal expected reward (Ardis & Brown, 2009). However, this is only guaranteed in the special case of state-transition dynamics that are deterministic. We investigate the discrepancy for inference in stochastic dynamics in more detail in Section 6.3 where we provide further insight into the underlying reason for this difference.

3.2 Optimal Control Approaches

Early control approaches to imitation learning directly model the (presumably near-optimal) policy for future execution in similar situations. Perhaps the most successful application of this approach is ALVINN (Pomerleau, 1989), a neural network model that learns a direct mapping from camera input to vehicle control for autonomous vehicle navigation on roadways. This approach is well-suited for predicting immediate stimulus-response behavior, like keeping a vehicle on the roadway. However, it is more difficult for predicting purposeful long-term decision making, like planning a route between two cities. More recent approaches, including that of this thesis, attempt to learn an underlying reward function to explain goal-directed, adaptive behavior rather than directly learning the policy.

3.2.1 Inverse Optimal Control

Inverse optimal control (Boyd et al., 1994; Ng & Russell, 2000), originally posed by Kalman (1964), describes the problem of recovering an agent's reward function, R(s,a), given a controller or policy, when the remainder of the PRMDP, $\mathcal{M}/R$, is known. We assume the PRMDP setting (Definition 2.10). Vectors of reward factors $\mathbf{f}_{s,a}$ describe each available action, and the reward function is assumed to be a linear function of those factors, $R(s,a) = \theta^\top \mathbf{f}_{s,a}$ parameterized by reward weights, θ. Ng & Russell (2000) formulate inverse optimal control as the recovery of reward weights, θ, that make demonstrated behavior optimal. Unfortunately this formulation is ill-posed in general.

Remark 3.1 (Ng & Russell (2000))**.** *Demonstrated behavior may be optimal for many different reward weights, including degeneracies (e.g., the vector of all zeros).*

Ng & Russell (2000) propose a heuristic for preferring solutions that discourage deviation from demonstrated behavior and simple rewards (small L_1 norm). However, many other heuristics can be employed and the justification for choosing this particular heuristic is unclear. Chajewska et al. (2001) maintain a Bayesian distribution over the utility weights that are consistent with demonstrated behavior being optimal. They employ a prior that discourages degenerate weights. An expensive Markov chain Monte Carlo sampling procedure is required to maintain this distribution. For both of these approaches, a very serious problem still remains when modeling demonstrated behavior.

Remark 3.2. *There may be no feasible utility weight apart from degeneracies for which demonstrated behavior is optimal.*

Demonstrated behavior may be inherently sub-optimal for a variety of reasons (discussed in Section 1.4) due to the fact that the Markov decision process is only an approximation of reality and decisions may be based on information that is difficult to observe or model. Allowing sub-optimality and uncertainty of behavior in the model of demonstrated behavior is needed to address this problem. Models based on the assumption of optimality without robustness to noisy (*i.e.*, sub-optimal) behavior are inappropriate as a result (Remark 3.2).

3.2.2 Feature Matching Optimal Policy Mixtures

Abbeel & Ng (2004) propose recovering reward weights so that a planner based on those reward weights and the demonstrated trajectories have equal reward (in expectation) for any choice of parameters. This formulation reduces to matching the planner and demonstrated trajectories' expected **feature counts**, $\mathbf{f}_\zeta = \sum_{s,a \in \zeta} \mathbf{f}_{s,a}$:

$$\sum_\zeta \mathrm{P}_{\text{planner}}(\zeta)\mathbf{f}_\zeta = \sum_\zeta \tilde{\mathrm{P}}(\zeta)\mathbf{f}_\zeta. \tag{3.3}$$

This feature-matching constraint guarantees that the model's expected performance will be equivalent to the demonstrated behavior's realized performance for any unknown linear reward function parameterized by θ^*, as

$$\sum_{\zeta} \mathrm{P}_{\text{planner}}(\zeta)\ ({\theta^*}^\top \mathbf{f}_\zeta) = ({\theta^*}^\top \mathbf{f}_{\tilde{\zeta}}). \tag{3.4}$$

Abbeel & Ng (2004) employ a series of deterministic policies obtained from "solving" the optimal MDP for the distribution over trajectories. When sub-optimal behavior is demonstrated (due to the agent's imperfection or unobserved reward factors), mixtures of those optimal policies are required to match feature counts:

$$\mathrm{E}_{\text{planner}}[\mathbf{f}] = \sum_i \mathrm{P}(\theta_i) \left(\sum_{\zeta} \mathrm{P}(\zeta|\pi^*(s|\theta_i))\ \mathbf{f}_\zeta \right). \tag{3.5}$$

The proposed approach for generating a series of candidate policies is shown as Algorithm 3.1

Algorithm 3.1 Policy mixture learning algorithm

Require: State-based features $\mathbf{f}$, Example MDP $\mathcal{M}$, Example feature counts $\tilde{\mathrm{E}}[\mathbf{f}]$,
Ensure: A sequences of policies $\{\pi^{(i)}\}_{i=0}^N$ such that the empirical feature counts, $\tilde{\mathrm{E}}[f]$, are matched by a mixture of $\{\pi^{(i)}\}_{i=0}^N$ within ϵ.

1: Randomly pick $\pi^{(0)}$ and compute $\mathrm{E}[\mathbf{f}|\pi^{(0)}]$.
2: **while** for $i = 1$ to ∞ **do**
3: Compute $t^{(i)} = \max_{\theta:||\theta||_2=1} \min_{j\in\{0..(i-1)\}} \theta^\top(\tilde{E}[\mathbf{f}] - \mathrm{E}[\mathbf{f}|\pi^{(j)}])$ and let $\theta^{(i)}$ be the parameter values obtained by this maximum
4: **if** $t^{(i)} \leq \epsilon$ **then**
5: **return** $\{\pi^{(i)}\}_{i=0}^N$
6: **end if**
7: Compute optimal policy $\pi^{(i)}$ for $\theta^{(i)}$ in MDP $\mathcal{M}$
8: Compute $\mathrm{E}[\mathbf{f}|\pi^{(i)}]$ in MDP $\mathcal{M}$
9: **end while**

The algorithm returns a series of policies that define a convex hull of expected feature counts that contains the demonstrated feature counts. Proposed selection criteria for a single policy from those returned either involve expert intervention (and evaluation) or discard the performance guarantees of the approach relative to demonstrated behavior. Alternately, a mixture of policies can be obtained from the convex hull that will match feature counts.

While this approach resolves some of the ambiguity of the original inverse optimal control problem (Remarks 3.1 and 3.2), it does not resolve it entirely. Many mixtures of policies that match feature counts can be obtained from the set of policies returned by the algorithm and from other policies obtained using different randomly-chosen initial policies for the algorithm. Additionally,

in the special case that only a single policy rather than a policy mixture is required to match feature counts, there are still many choices of parameters weights that will realize that policy.

Remark 3.3. *When demonstrated behavior is optimal for a single choice of reward parameters, infinitely many other choices of parameters will also make it optimal.*

While Remark 3.3 is trivially true due to the invariance of the optimal policy to positively scaling reward parameters, it is more generally also true since a policy is optimal within some polytope of utility parameters in the reward parameter simplex and often not for only a single point.

Even ignoring these unresolved ambiguities, mixing optimal policies has undesirable properties for any selection criteria. The resulting behavior obtained by first sampling parameters, θ_i, and then the optimal policy in Equation 3.5 varies drastically depending on which parameters are sampled. This can be alleviated by replacing the mixture of optimal policies with a stationary stochastic policy as suggested by Syed et al. (2008) (following Theorem 2.8). However, whether mixing optimal policies or employing a corresponding stochastic policy, the model can still have very poor predictive capabilities.

Remark 3.4. *A learned policy mixture of optimal policies (or the corresponding single stochastic policy) can assign zero probability for demonstrated behavior if that behavior is sub-optimal for any choice of cost weights.*

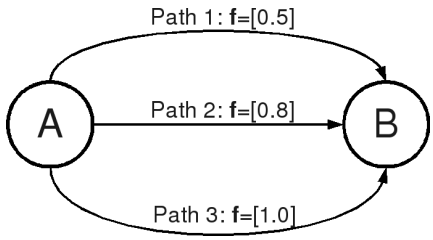

Figure 3.4: A simple example with three action choices where the mixture of optimal policies can have zero probability for demonstrated behavior if path 2 is demonstrated.

Consider the simplified example in Figure 3.4 with three action choices corresponding to Paths 1, 2, and 3, and one feature characterizing each action choice. For sake of argument, assume that demonstrated behavior consists solely of selecting path 2. There are only two non-degenerate optimal policies: for π_1, $\theta < 0$ implies path 1 should be selected, and for π_2, $\theta_2 > 0$ implies path 3 should be selected[1]. These can be appropriately mixed with $\alpha_1 = 0.4$ and $\alpha_2 = 0.6$ so that demonstrated feature counts and the feature counts expected under the optimal policy mixture match, however note that there is zero probability of the demonstrated action (path 2) in this solution.

[1]For parameters $\theta = 0$, ties are typically broken according to an arbitrarily chosen secondary tie-breaking criteria. For example, if the order of consideration were employed, Path 1 would be selected for $\theta = 0$.

3.2.3 Maximum Margin Planning

Ratliff et al. (2006) resolve the ambiguity of choosing a single set of parameter weights to make demonstrated behavior optimal (Remark 3.3) by posing inverse optimal control as a maximum margin problem. They reduce a quadratic program for solving this problem into the following convex objective:

$$C_q(\theta) = \frac{1}{n}\sum_{i=1}^{n} \beta_i \left(\max_{\mu \in \mathcal{G}_i} \left(\theta^\top F_i + l_i^\top\right) \mu - \theta^\top F_i \mu_i \right)^q + \frac{\lambda}{2}||\theta||^2, \tag{3.6}$$

where μ are expected state visitation counts constrained by the MDP's structure ($\mathcal{G}$), μ_i are the empirical visitation counts for example i, and l_i is an augmented state loss, $l_i : S \to \mathbb{R}$, that assigns higher loss to states visited in example i.

Algorithm 3.2 Maximum margin planning

Require: State-based features $\mathbf{f}$, Example MDPs $\{\mathcal{M}_i\}$, Example feature counts $\{\tilde{\mathrm{E}}[\mathbf{f}]_i\}$, Example loss augmentation $\{l_i\}$, Learning rate r, Regularization parameter C, Iterations T
Ensure: Parameters, θ, that (approximately) optimize the maximum margin objective.
1: $\theta \leftarrow 0$
2: **for** $t = 1$ to T **do**
3: Compute optimal $\pi^*(s)$ for each loss-augmented cost map $\theta^\top \mathbf{f}_i + l_i^\top$ in MDP $\mathcal{M}_i$
4: Compute $\mathrm{E}[\mathbf{f}|\pi^*(s)]_i$ in MDP $\mathcal{M}_i$
5: Compute sub-gradient $g = \theta + C(\sum_i (\mathrm{E}[\mathbf{f}|\pi^*(s)]_i - \tilde{\mathrm{E}}[\mathbf{f}]_i))$
6: $\theta \leftarrow \theta - \frac{r}{t} g$
7: **end for**
8: **return** θ

The sub-gradient algorithm (Algorithm 3.2) for optimizing the maximum margin planning objective (Equation 3.6) operates by effectively lowering the costs of actions that are away from the demonstrated trajectory and finding cost weights making the cost-augmented demonstrated trajectory optimal. While the approach yields a unique solution, it suffers from significant drawbacks when no single reward function makes demonstrated behavior both optimal and significantly better than any alternate behavior. This arises quite frequently when, for instance, the behavior demonstrated by the agent is imperfect, or the planning algorithm only captures a part of the relevant state-space and cannot perfectly describe the observed behavior.

A feature boosting approach (Ratliff et al., 2007) partially addresses this sub-optimal behavior problem by creating a larger (non-linear) class of features to explain demonstrated behavior as optimal. While this may succeed, it also runs the risk of crafting features that overfit to the training data. The methods presented in this thesis differ from maximum margin planning in the explicit treatment of uncertainty under the maximum causal entropy approach. This uncertainty is needed to address suboptimal behavior.

3.2.4 Game-Theoretic Criteria

Syed & Schapire (2007) employ a game-theoretic approach that maximizes performance relative to the demonstrated behavior for the worst-case choice of reward function. When background knowledge is available indicating the sign of the different cost factors (*i.e.*, whether each feature is "good" and should be maximized or "bad" and should be minimized), the learned model can outperform demonstrated behavior.

Remark 3.5 (Syed & Schapire (2007))**.** *When it is assumed that* $\forall_i\ \theta_i \geq 0$*, then if* $\forall_i\ \mathrm{E}_P[\boldsymbol{f}_i] \geq \mathrm{E}_{\tilde{\mathrm{P}}}[\boldsymbol{f}_i]$*, the distribution* P *has expected reward guaranteed to be higher than the reward of demonstrated behavior on the unknown cost function* θ^**.*

This approach similarly yields a mixture of policies, which is later converted to a stationary policy using Theorem 2.8 and learned using a linear programming formulation (Syed et al., 2008). It also suffers from potentially assigning zero probability to demonstrated behavior.

3.2.5 Boltzmann Optimal-Action-Value Distribution

An alternative to mixing optimal policies to obtain a non-deterministic behavior model is to directly obtain a distribution as a function of the MDP's optimal action values, $Q_\theta(s,a)^*$, obtained by first solving the Bellman equation (Theorem 2.7). A natural approach to regressing from a set of real values to a probability distribution is the **Boltzmann probability distribution**. Inference under this model was first proposed for goal inference (Baker et al., 2006, 2007) using the true costs of generated behavior in synthetic experiments:

$$\mathrm{P}(\text{action } a|\text{state } s) = \frac{e^{\beta Q_\theta(s,a)}}{\sum_{\text{action } a'} e^{\beta Q_\theta(s,a')}}. \tag{3.7}$$

The scale parameter, β, determines the stochasticity of the policy. As $\beta \to \infty$, the policy becomes deterministic[2]. When $Q_\theta(s,a)$ is learned from data, its scale is also learned and β is superfluous.

Two approaches for learning parameters in this model have been proposed. Neu & Szepesvári (2007) employ this distribution within a loss function penalizing the squared difference in probability between the model's action distribution and the demonstrated action distribution. Ramachandran & Amir (2007) utilize it within a Bayesian approach to obtain a posterior distribution over reward values using Markov Chain Monte Carlo simulation. The main weaknesses of the model are summarized by Remark 3.6 and Remark 3.7.

Remark 3.6 (Neu & Szepesvári (2007))**.** *The likelihood of data in the Boltzmann optimal-action-value distribution is not convex in its parameters* θ*.*

[2]This assumes that no two actions have equivalent reward. If there are multiple optimal actions, they will have equal probability.

The non-convexity of the data likelihood function under the Boltzmann model (Remark 3.6) makes finding optimal model parameters difficult, and non-convex distributions such as this are as a result generally less preferred than convex models for computational efficiency reasons.

Remark 3.7. *The relationship between a policy's expected reward and its probability in the Boltzmann optimal-action-value distribution is not monotonic.*

Additionally, the non-monotonic relationship between a policy's expected reward and its probability within the model (Remark 3.7) has a few undesirable consequences: the most likely policy within the model does not necessarily have the highest probability within the model and policies with equal reward do not necessarily have equal probability (and vice versa). Also, this model provides no guarantees of performance relative to the demonstrated behavior.

3.2.6 Robust and Approximate Optimal Control

Another branch of research makes optimal decisions based on the Kullback-Leibler divergence of decision setting dynamics to realize robust control techniques (Bagnell, 2004; Nilim & El Ghaoui, 2005). Specifically, this technique provides strong PAC-style performance guarantees when the transition dynamics of a decision setting are only approximately known. It was recently observed that the resulting probabilistic model is a linearization of the Bellman equations (Kappen, 2005; Todorov, 2006). This linearized view enables more efficient techniques for approximating the optimal policy's value by solving a system of linear equations. The problem is formulated as a continuous control problem where the action cost is defined as the Kullback-Liebler divergence of the employed action's implied dynamics and a set of "passive" dynamics, $P_0(s_j|s_i)$. The actions in this approach directly determine the next state—there are no stochastic dynamics given the action, but a high cost is paid to employ a deterministic policy that differs from the passive dynamics. The optimal action under this formulation yields stochastic dynamics:

$$P^*_{\mathrm{KL}}(s_j|s_i) = \frac{\mathrm{P}_0(s_j|s_i)\ e^{v^*_{\mathrm{KL}}(s_j)}}{\sum_{s_k} P_0(s_k|s_i)\ e^{v^*_{\mathrm{KL}}(s_k)}}. \tag{3.8}$$

In general, state transitions in Markov decision processes are dependent on the action and each action does not deterministically lead to a different next state. Todorov proposes a method for embedding general Markov decision processes within this framework (2009b), but this approach is approximate and only anecdotal evidence has been provided to demonstrate its validity.

The compositionality of control laws within the linear Bellman approach is described by Todorov (2009a) and employed (approximately) for generating realistic character animation (da Silva et al., 2009).

Remark 3.8 (following Todorov (2009a))**.** *Let $\pi^{(1)}_{KL}$ and $\pi^{(2)}_{KL}$ be optimal KL policies for terminal reward $V^{(1)}_{final}(s)$ and $V^{(2)}_{final}(s)$ in a first-exit linear Bellman model with corresponding path integrals*

$Z_{V^{(1)}}(s)$ and $Z_{V^{(2)}}(s)$. Then for $\alpha \in [0,1]$, $\beta\pi^{(1)}_{KL} + (1-\beta)\pi^{(2)}_{KL}$ is an optimal linear Bellman policy for terminal reward $\alpha V^{(1)}_{final}(s) + (1-\alpha)V^{(2)}_{final}(s)$, where $\beta = \frac{\alpha z_{V^{(1)}}(s)}{\alpha Z_{V^{(1)}}(s) + (1-\alpha) Z_{V^{(2)}}(s)}$.

Our work generalizes this idea of efficient composability to a full Bayesian treatment of latent goal states (Ziebart et al., 2008c) in Chapter 11. Specifically, it not only incorporates arbitrary terminal rewards, but also addresses appropriately updating those rewards based on observed behavior when the goal of the behavior is unknown. We apply this approach to vehicle route preference prediction in Chapter 12 and pedestrian trajectory prediction in Chapter 13.

A learning approach under the linear Bellman equation perspective (Krishnamurthy & Todorov, 2010) is identical to the maximum causal entropy model of this thesis for deterministic dynamics (Ziebart et al., 2008b), except that a value function rather than a (feature-based) cost function is learned[3]. Value function learning can be viewed as a special case of cost function learning where each state has its own unique feature. The advantage of learning the value function is the simplicity of not having to solve a forward inference problem. For some domains, that forward inference problem can be prohibitively expensive. However, there is also a large disadvantage because the resulting model is then unable to generalize. For example, changes in end-state constraint (*i.e.*, goal), state transition dynamics, or transfer to similarly-parameterized MDPs require a learned cost function rather than a learned value function. Krishnamurthy & Todorov (2010) propose a linear regression approach to obtain a parametric cost function from the learned values, however this approach provides no predictive guarantees and is not evaluated in that work. A method of generalizing by estimating the value function as a mixture of Gaussians is proposed to reduce dependence on training data for accurate estimation, but this still does not transfer to other goals or settings.

We investigate some of the restrictions of linearizing the Bellman equation relative to the principle of maximum causal entropy approach developed in this thesis in Section 6.3 and relate the compositionality of optimal control in the linearized Bellman equation approach to our techniques for efficient Bayesian inference of latent goals in Chapter 11.

3.2.7 Noise-Augmented Optimal Actions

A final, and quite general approach from the optimal control perspective for making an optimal policy stochastic is to augment the optimal value function with a sampled noise distribution, $\epsilon_{s,a}$, and generate a stochastic policy using many sampled deterministic policies:

$$\pi(s) = \operatorname*{argmax}_a Q^*(s,a) + \epsilon_{s,a}. \tag{3.9}$$

[3]The claimed "passive dynamics" generalization is equivalent to optimizing the relative entropy and is generalized via Remark 5.3 by simply adding corresponding action features.

Any distribution can be employed for the error term, however the generality comes at a cost: simulation is required rather than closed-form analysis. Nielsen & Jensen (2004) take this approach to model decisions in an influence diagram and to then learn the underlying utility function upon which $Q^*(s, a)$ is based using Markov chain Monte Carlo sampling strategies. In this thesis, we employ basic principles to obtain a theoretically-justified error term. This provides predictive guarantees and computational efficiency benefits that the noise-augmented approach generally lacks.

3.3 Discrete Choice Theory

Economists have long studied models for estimating the probability of different choice selections from a set of discrete options under various axiomatic formulations and noise assumptions. We review some of the approaches established from that perspective here.

3.3.1 Multinomial and Conditional Logit

Choice models estimate the probability that an individual with characteristics defined by vector $\mathbf{x}_i$ will select one of $c \in C$ choices (*i.e.*, $\mathrm{P}(y = c|\mathbf{x}, C)$). Luce (1959) derived an axiomatic probabilistic model based on the **independence of irrelevant alternatives** (IIA) axiom. Under the IIA axiom, the relative odds between two choices are unaffected by the addition of options to the set of choices.

$$\forall_{c_1,c_2 \in C_a,C_b} \frac{\mathrm{P}(y = c_1|\mathbf{x}, C_a)}{\mathrm{P}(y = c_2|\mathbf{x}, C_b)} = \frac{\mathrm{P}(y = c_1|\mathbf{x}, C_a)}{\mathrm{P}(y = c_2|\mathbf{x}, C_b)} \tag{3.10}$$

The resulting **multinomial logit** probabilistic model has the following form:

$$\mathrm{P}(y = c|\mathbf{x}, C_a, \theta) = \frac{e^{\theta_c^\top \mathbf{x}}}{\sum_{c' \in C} e^{\theta_{c'}^\top \mathbf{x}}}, \tag{3.11}$$

which is a log-linear (logistic regression) model.

McFadden (1974) proposes employing characteristics of the available options, $\mathbf{z}_c$, within a model based on the IIA axiom. When both characteristics of the individual and the options are employed, the probability distribution of this **conditional logit model** takes the following form:

$$\mathrm{P}(y = c|\mathbf{x}, \mathbf{z}, C, \theta, \phi) = \frac{e^{\theta_c^\top \mathbf{x} + \phi^\top \mathbf{z}_c}}{\sum_{c' \in C} e^{\theta_{c'}^\top \mathbf{x} + \phi^\top \mathbf{z}_{c'}}}, \tag{3.12}$$

which is similarly a log-linear model. This approach has the advantage of providing specification to the predictive probability distribution based on the characteristics of the different available options.

3.3.2 Criticisms of the Independence of Irrelevant Alternatives Axiom

A common criticism of the IIA axiom (and the multinomial and conditional logit models that employ it) is its inability to take into account (near) perfect substitutes. Consider the "Red Bus/Blue Bus" example (McFadden, 1974), where the options of driving a car or taking a red bus are available to a commuter with equal probability under the model. If a blue bus option is added, and commuters are indifferent to bus color, one would expect a probability of 50%, 25%, and 25% for car, red bus, and blue bus, but instead each is equally probable under the IIA axiom.

A number of alternate discrete choice models have been proposed that relax the IIA axiom to address this scenario.

3.3.3 Nested Logit

The *nested logit* discrete choice model (Ben-Akiva, 1974) incorporates correlation among similar groups of disjoint choices (nests), $C_1, C_2, ...$ such that $C = \cup_k \, C_k$ and $\forall_{j \neq k} \; C_j \cap C_k = \emptyset$. The probability of a choice i in clique C_k is specified by the chain rule:

$$\mathrm{P}(i) = \mathrm{P}(C_k)\,\mathrm{P}(i|C_k), \tag{3.13}$$

with probabilities defined by:

$$\begin{aligned} \mathrm{P}(C_k) &= \frac{e^{Z_{C_k} + \lambda \ln \sum_{i \in C_k} e^{V_i}}}{\sum_{j=1}^{n} e^{Z_{C_j} + \lambda \ln \sum_{i \in C_j} e^{V_i}}} \text{ and} \\ \mathrm{P}(i|C_k) &= \frac{e^{V_{C_i}}}{\sum_{j \in C_k} e^{V_{C_j}}}. \end{aligned} \tag{3.14}$$

The $\lambda \in [0, 1]$ parameter in this model is related to the correlation between choices within each of the nests. When $\lambda = 1$, this model reduces to a standard logit model.

3.4 Discussion

We now discuss the relations of the reviewed approaches to each other and to the maximum causal entropy approach contributed by this thesis.

3.4.1 Connections Between Approaches

Despite being developed under varying perspectives with differing assumptions and for different purposes, there are a number of similarities between these approaches. The linear Bellman optimal control model (Todorov, 2006) and the conditional multinomial logit discrete choice model

(McFadden, 1974) are both exponential family models that can be equivalently expressed as a special form of conditional random fields with potentials that correspond to rewards (Ardis & Brown, 2009; Toussaint, 2009). Evaluating the robust MDP model by the power method is equivalent to forward or backward algorithms for chain conditional random field inference and is a special case of the nested logit model's hierarchical inference procedure (Equation 3.14).

The independence of irrelevant alternatives axiom of discrete choice theory also has analogies in the optimal control approaches for decision modeling. Specifically, the correlation-based parameter of the nested logit model has a number of parallels. The Boltzmann optimal-action-value distribution does not obey the IIA axiom and is similar to the $\lambda = 0$ nested logit model. More generally, the discount factor, which is commonly employed to make the Bellman equations for infinite time horizons convergent, when applied within the linear Bellman optimal control model is very similar to the effect that λ has in the nested logit model.

3.4.2 Relation to Thesis Contributions

Information revelation settings

The primary contribution of this thesis is a principled approach for modeling the conditional probabilities of variables in settings with side information that is revealed over time from a known conditional probability distribution. Markov decision processes with stochastic dynamics match this information revelation setting because the next state is random and only revealed *after* an action is made. Conditional random fields are ill-suited for this setting because they are based on the principle of maximum conditional entropy, which assumes full knowledge of random side information in advance. As a consequence, there is a disconnect between maximizing conditional likelihoods in a conditional random field and maximizing the expected reward in a corresponding MDP (Remark 6.17). Similarly, the linear Bellman equation model relies on an imprecise embedding to model general MDPs because linearization alone is inappropriate for dealing with the information availability of the general MDP.

The maximum causal entropy approach of this thesis provides a means for probabilistically modeling behavior within settings characterized by sequentially revealed side information (*e.g.*, such as the sequence of states in a Markov decision process) so that maximum likelihood policy estimation is equivalent to expected reward maximization. This is in contrast to the Boltzmann optimal value action model, where the maximum likelihood policy and the maximum expected reward policy within the Markov decision process do not necessarily match. An additional advantage of the maximum causal entropy approach is that the optimization for learning parameters in the model is convex.

Cost potential function learning

In the special case that the corresponding Markov decision process has deterministic state-transition dynamics, revealing the next state provides no new information. This is due to the fact that the next

state is already known from the choice of action and current state. In this setting, the principle of maximum causal entropy reduces to a probabilistic model that matches the linear Bellman model (Todorov, 2006), conditional random field (Ardis & Brown, 2009; Toussaint, 2009), and conditional logit (McFadden, 1974) models of behavior.

While the development of the linear Bellman model (Todorov, 2006) precedes the maximum entropy approach for decision modeling (Ziebart et al., 2008b), its purpose was for efficient approximate inference of the optimal policy and optimal value function. More recent attempts to apply the linear Bellman model for learning (Krishnamurthy & Todorov, 2010) propose learning value functions rather than the cost functions learned in the maximum entropy approach. As previously discussed, the computational benefits of the approach come at the cost of not being able to generalizes to changes in the problem setting.

Similarly, the conditional random field model of Markov decision processes has only been previously applied for purposes of inference based on known reward values (Ardis & Brown, 2009; Toussaint, 2009). We note that our work (Ziebart et al., 2008b) represents the first employment of this model for the purpose of learning from observed behavior. Additionally, the use of features associated with "states" rather than observations is atypical from the standard CRF perspective. The principle of maximum causal entropy generalizes the validity of this approach to settings with stochastic state dynamics.

The conditional logit model is perhaps the most similar to the maximum entropy model of decision making on decision tasks with no information revelation. It is important to note the limits of the conditional logit approach in its previous applications. For route preference modeling, for instance, the first step of applying logit models when given a road network with infinite possible paths between two points is to employ a route selection algorithm that generates a tractable subset of paths to consider. We view our approach as an extension of the efficiency of the conditional logit model from trees of decisions to graphs with (potentially) infinite sequences of choices. Maximum causal entropy also generalizes the discrete choice models to settings with stochastic state transition dynamics.

Part II

Theory

Overview of Part II

Part II of the thesis provides the formulations and theoretical properties of the maximum causal entropy approach needed for modeling sequences of behavior, as well as the justifications for its use as a predictive tool.

Chapter 4 reviews the theory of causal information. This less widely known extension of information theory enables its applicability to settings with feedback and interaction.

Chapter 5 presents the general formulation for modeling a sequence of variables that is causally conditioned on a sequence of additional sequentially-revealed variables. Two types of constraints are introduced: affine equality constraints and convex inequality constraints, which generally correspond to efficiency and rationality requirements in this thesis.

Chapter 6 applies the general maximum causal entropy formulation to the inverse optimal control setting of finding the reward, utility, or cost function of the parametric Markov decision process or linear-quadratic regulator frameworks introduced in Chapter 2. Equality constraints are employed to match statistical expectations of characteristics of demonstrated behavior. Inference under this model corresponds to a softened application of the Bellman equation, leading to a number of simple inference algorithms in Part III.

Chapter 7 extends the maximum causal entropy approach to the influence diagram framework (Chapter 2). Whereas in the inverse optimal control setting of Chapter 6, only future side information is unknown or latent at each point in time, the formulation of this chapter allows current and past information to also be unknown. This corresponds to common decision settings with partial information, such as the POMDP, and to collaborative multi-agent coordination. We establish an important boundary on the applicability of the maximum causal entropy approach: perfect recall of past decisions is required for many of the theoretical niceties introduced in Chapter 5 to hold.

Chapter 8 demonstrates the wider applicability of the maximum causal entropy approach by applying it to multi-agent game play. The inequality constraints of Chapter 5 are employed to enforce rationality requirements of players' strategies in sequential games. This enables a novel, well-specified correlated equilibrium solution concept with log-loss prediction guarantees in the sequential setting of Markov games.

Together, the theoretical developments of these five chapters provide a broad scope of applicability for the maximum causal entropy approach for predicting behavior under a number of behavior frameworks.

Chapter 4

The Theory of Causal Information

"Nobody knows what entropy really is, so in a debate you will always have the advantage."
– John von Neumann (Mathematician, 1903–1957)

The field of information theory was developed primarily by Shannon's seminal communications work (1948). It enables the quantification of the amount of information provided by random variables. Information measures for joint, conditional, marginal, and relative distributions are well known. They are frequently employed in communication, control, and machine learning applications. In this chapter, we review those measures and their properties. We then review the less widely known causal extension of information theory. This extension enables the applicability of information theory in settings of interaction and feedback. It serves as the backbone of our approach for creating predictive distributions for human behavior in those settings.

4.1 Information Theory

We provide a brief overview of information-theoretic measures. We refer the reader to Cover & Thomas (2006)'s excellent book on the topic for a more thorough coverage of information theory and its applications.

4.1.1 Entropy and Information

Shannon's **information entropy** of distribution P over random variable X measures the *uncertainty* of that probability distribution.

Definition 4.1. *The **entropy** of a discrete random variable, X, distributed according to* P *is:*

$$\begin{aligned} H_{\mathrm{P}}(X) &= \mathrm{E}_{\mathrm{P}(X)}[-\log \mathrm{P}(x)] \\ &= -\sum_{x \in \mathcal{X}} \mathrm{P}(x) \log \mathrm{P}(x). \end{aligned} \tag{4.1}$$

For continuous-valued variables[1]*, the entropy is:*

$$H_{\mathrm{P}}(X) = \int_x \mathrm{P}(x) \log \mathrm{P}(x)\, dx. \tag{4.2}$$

When $\log_2$ is used, the discrete entropy (4.1) measures the average number of bits needed for a binary message that encodes a sample from distribution P. A uniform distribution maximizes this entropy measure, while an entropy of zero corresponds to a single point distribution.

When X is a vector of variables, $\mathbf{X}_{1:T}$, rather than a scalar, the **joint entropy** measures the uncertainty of the joint distribution, $P(\mathbf{X}_{1:T})$.

Definition 4.2. *The **joint entropy** of the vector of random variables, $\boldsymbol{X}_{1:T}$, distributed according to $\mathrm{P}(\boldsymbol{X}_{1:T})$ is:*

$$\begin{aligned} H_{\mathrm{P}}(\boldsymbol{X}_{1:T}) &= \mathrm{E}_{\mathrm{P}(\boldsymbol{X}_{1:T})}[-\log \mathrm{P}(\boldsymbol{x}_{1:T})] \\ &= -\sum_{\boldsymbol{x}_{1:T}} \mathrm{P}(\boldsymbol{x}_{1:T}) \log \mathrm{P}(\boldsymbol{x}_{1:T}). \end{aligned} \tag{4.3}$$

The **cross entropy** measures the amount of uncertainty of the distribution P under a different distribution Q both over random variable X.

Definition 4.3. *The **cross entropy** of distribution P under distribution Q is:*

$$\begin{aligned} H(\mathrm{P}, Q) &= \mathrm{E}_{\mathrm{P}(X)}[-\log Q(x)] \\ &= -\sum_{x \in \mathcal{X}} \mathrm{P}(x) \log Q(x). \end{aligned} \tag{4.4}$$

From the coding perspective, the cross entropy is the expected number of bits needed to encode samples from P under the incorrect probability distribution, Q. It is equivalent to the **log-loss** of distribution Q evaluated by predicting data actually distributed according to P.

The **conditional entropy** extends information theory to conditional probability distribution settings. It measures the average amount of information needed to recover Y given that X is known.

Definition 4.4. *The **conditional entropy** of Y given X distributed according to $\mathrm{P}(Y|X)$ is:*

$$\begin{aligned} H_{\mathrm{P}}(Y|X) &= \mathrm{E}_{\mathrm{P}(X,Y)}[-\log \mathrm{P}(y|x))] \\ &= -\sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} \mathrm{P}(y, x) \log \mathrm{P}(y|x), \end{aligned} \tag{4.5}$$

for scalar random variables Y and X and naturally extends to vectors of random variables.

[1]We will typically only consider the discrete-valued expansion in this chapter. The continuous-valued analog can be obtained from the continuous-valued definition of expectations.

The entropy of the previously described distributions are implicitly relative to a uniform distribution. In general, other relative distributions can be employed. The **relative entropy** measures the additional amount of information needed to code samples of P according to a different encoding distribution, Q.

Definition 4.5. *The **relative entropy** of distribution* P *relative to distribution* Q *is:*

$$\begin{aligned} H(\mathrm{P}\,//Q) &= E_{\mathrm{P}(X)}\left[\log \frac{\mathrm{P}(x)}{Q(x)}\right] \\ &= \sum_{x\in\mathcal{X}} \mathrm{P}(x) \log \frac{\mathrm{P}(x)}{Q(x)}. \end{aligned} \tag{4.6}$$

We employ this modified notation to avoid conflict with causal entropy measure notation[2].

The relative entropy is more commonly known as the **Kullback Leibler-divergence** (Kullback & Leibler, 1951) between P and Q: $D_{KL}(\mathrm{P}\,||Q) \triangleq H(\mathrm{P}\,//Q)$.

Definition 4.6. *The **mutual information** of two variables distributed according to* $\mathrm{P}(Y, X)$ *is:*

$$\begin{aligned} I(X;Y) &= \mathrm{E}_{\mathrm{P}(X,Y)}\left[\log\left(\frac{\mathrm{P}(x,y)}{\mathrm{P}(x)\,\mathrm{P}(y)}\right)\right] \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \mathrm{P}(x,y)\log\left(\frac{\mathrm{P}(x,y)}{\mathrm{P}(x)\,\mathrm{P}(y)}\right). \end{aligned}$$

The **mutual information** (Definition 4.6) provides the amount of information shared by two random variables.

Definition 4.7. *The **conditional mutual information** of two variables conditionally distributed (given an additional variable) according to* $\mathrm{P}(Y, X|Z)$ *is:*

$$\begin{aligned} I(X;Y|Z) &= \mathrm{E}_{P(X,Y,Z)}\left[\log\left(\frac{\mathrm{P}(x,y|z)}{\mathrm{P}(x|z)\,\mathrm{P}(y|z)}\right)\right] \\ &= \sum_{z\in\mathcal{Z}} P(z) \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x,y|z)\log\left(\frac{\mathrm{P}(x,y|z)}{\mathrm{P}(x|z)\,\mathrm{P}(y|z)}\right) \end{aligned}$$

Lastly, the **conditional mutual information** (Definition 4.7) provides the amount of information shared by two random variables conditioned on a third variable.

[2]Typically the relative entropy is denoted as $H(\mathrm{P}\,||Q)$. However, this conflicts with the existing notation employed for the causal entropy we will employ later. Hence, we employ the alternate notation

4.1.2 Properties

There are many well-known and useful identities relating the previously introduced information-theoretic measures. We list some of them here.

Remark 4.8. *A **chain rule of entropies** relates the joint and conditional entropies:*

$$H(\boldsymbol{X}_{1:T}) = \sum_t H(X_{t+1}|\boldsymbol{X}_{1:t}).$$

Remark 4.9. *Following Remark 4.8, the conditional entropy can be expressed as:*

$$H(Y|X) = H(Y,X) - H(X).$$

Remark 4.10. *The entropy, relative entropy, and cross entropy are related by:*

$$H(\mathrm{P}\,//Q) \triangleq H(\mathrm{P},Q) - H_{\mathrm{P}}(X).$$

Remark 4.11. *The mutual information can be defined in terms of the joint, conditional, and marginal entropies in a number of ways:*

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(X,Y) - H(X|Y) - H(Y|X). \end{aligned}$$

Remark 4.12. *A **chain rule of mutual information** provides the following relation between mutual information and conditional mutual information:*

$$\begin{aligned} I(\boldsymbol{X}_{1:T};\boldsymbol{Y}_{1:T}) &= \sum_t I(\boldsymbol{Y}_{1:T};X_t|\boldsymbol{X}_{1:t-1}) \\ &= \sum_t I(\boldsymbol{X}_{1:T};Y_t|\boldsymbol{Y}_{1:t-1}). \end{aligned}$$

We also list useful inequalities that relate those information-theoretic measures. These inequalities can be obtained using **Jensen's inequality**: $f(E[X]) \leq E[f(X)]$ for convex f.

Remark 4.13. *The entropy is always positive:* $H(X) \geq 0$.

Remark 4.14. *Conditioning on additional variables never increases entropy:*

$$\forall_{X,Y,Z}\ H_{\mathrm{P}}(Y|X) \geq H_{\mathrm{P}}(Y|X,Z)$$

with equality when Y is conditionally independent of Z given X.

Remark 4.15. *Gibbs' inequality assures that the KL divergence is always positive:*

$$D_{KL}(P||Q) \geq 0 \tag{4.7}$$

with equality if and only if $P = Q$.

4.1.3 Information Theory and Gambling

Information theory and gambling have many close connections. We summarize the main results of Cover & Thomas (2006) in this section to illustrate those connections. Many perspectives on gambling have focused on maximizing the expected payoff of a (set of) bet(s). This strategy has the unfortunate problem that over a large enough time span, the realized payoff by employing an expected utility maximizing strategy will be zero with high probability. This is known as the gambler's paradox. An alternate perspective is to maximize the **rate of growth** of investment.

More concretely, Cover & Thomas (2006) consider an m horse race where the payoff on a bet is $b_i o_i$, where b_i is the fraction of total wealth bet that horse i will win, and o_i is the payoff multiplier if horse i does win. It is assumed that the initial wealth of the gambler is 1, and all of the gambler's wealth is bet during each race. After a sequence of races, the gambler's wealth will be:

$$S_n = \prod_{t=1}^{n} S(Y_t), \tag{4.8}$$

where $S(Y) = b(Y)\ o(Y)$ and Y_t is the winning horse. The expected doubling rate of investment is then: $W(b,p) = E[\log S(Y)] = \sum_{k=1}^{m} p_k \log(b_k o_k)$ when the winning horse is distributed according to **p**.

Theorem 4.16 (from Cover & Thomas (2006))**.** *Proportional gambling, $\boldsymbol{b}^* = \boldsymbol{p}$, is log-optimal with optimal doubling rate:*

$$W^*(\boldsymbol{p}) = \sum_{i=1}^{k} p_i \log o_i - H(\boldsymbol{p}). \tag{4.9}$$

Theorem 4.17 (from Cover & Thomas (2006))**.** *When side information, X_t, is provided before each race, the optimal doubling rate,*

$$W^*(\boldsymbol{Y}|\boldsymbol{X}) = \sum_{i=1}^{k} \mathrm{P}(y|x) \log o_i - H(\boldsymbol{Y}|\boldsymbol{X}), \tag{4.10}$$

is obtained by $b^* = p(y|x)$. The difference in growth rate by having the side information is the mutual information between the side information and the winning horse variables:

$$\Delta W = W^*(\boldsymbol{Y}|\boldsymbol{X}) - W^*(\boldsymbol{Y}) = I(\boldsymbol{Y};\boldsymbol{X}). \tag{4.11}$$

Thus, there are very strong connections between information-theoretic measures and investment growth rates in gambling. The causal information investigated in this thesis expand this gambling perspective to settings where the outcomes of consecutive horse races are dependent, and the the revealed side information is also dependent on previous side information and race outcomes.

4.2 Causal Information

4.2.1 Sequential Information Revelation and Causal Influence

Many important problem settings with sequential data involve partial information, feedback, and interaction. For example, in decision settings with stochastic dynamics, future side information (*i.e.*, the future state), $\mathbf{X}$, is unknown during earlier points of interaction (Figure 4.1). As a consequence, the future side information should have no *causal* influence in our statistical models of interaction until after it is revealed. In other words, the specific instantiation value, x_{t_2}, should have no causal influence on y_{t_1} for $t_1 < t_2$. However, this does not imply a statistical independence—the influence of earlier interactions on future side information should not be ignored. In the case of stochastic dynamics, the actions chosen do influence the future state.

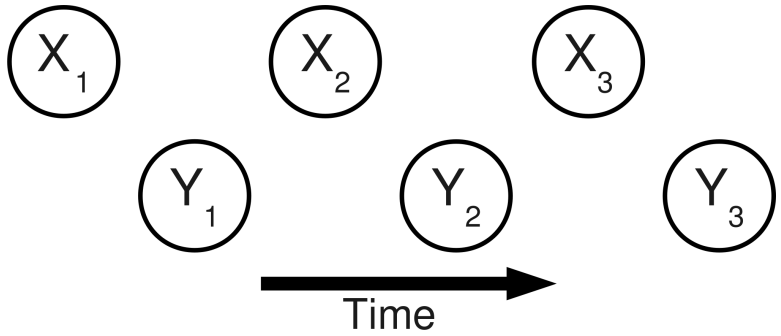

Figure 4.1: The sequence of side information variables, $\mathbf{X}$, and conditioned variables, $\mathbf{Y}$, that are revealed and selected over time.

More formally this notion of causal influence means that if a future side information variable were secretly fixed to some value by **intervention** (Pearl, 2000) rather than generated according to its conditional probability distribution, the distribution over all earlier variables would be unaffected by this change. In contrast, if by intervention some side information variable were fixed to some value, the value of variables *does* influence future conditioned variables.

4.2.2 Causally Conditioned Probability

The **causally conditioned probability** (Kramer, 1998) from the Marko-Massey theory of directed information (Marko, 1973; Massey, 1990) is a natural extension of the conditional probability, $\mathrm{P}(\mathbf{Y}|\mathbf{X})$, to the situation where each Y_t (for $t = 1..T$) is conditioned on only a portion of the $\mathbf{X}$ variables, $\mathbf{X}_{1:t}$, rather than the entirety, $\mathbf{X}_{1:T}$.

Definition 4.18. *Following the previously developed notation (Kramer, 1998), the probability of* $\mathbf{Y}$

***causally conditioned** on **X** is:*

$$\mathrm{P}(\boldsymbol{Y}^T||\boldsymbol{X}^T) \triangleq \prod_{t=1}^{T} \mathrm{P}(Y_t|\boldsymbol{X}_{1:t}, \boldsymbol{Y}_{1:t-1}). \tag{4.12}$$

We will often consider $\mathbf{Y}$ to be actions or controls applied to a system and $\mathbf{X}$ to be the system's state in this thesis. This interpretation can be useful in understanding the approach, but it is important to note that $\mathbf{Y}$ and $\mathbf{X}$ can represent any sequentially revealed information where the causal assumption that future side information does not influence earlier variables is reasonable.

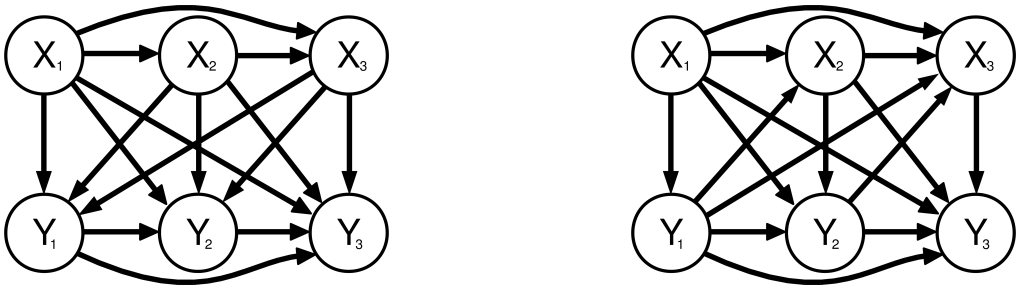

Figure 4.2: A sequence of variables, $\mathbf{Y}_{1:T}$ conditioned (left) and causally conditioned (right) on a corresponding sequence of side information variables, $\mathbf{X}_{1:T}$. In the conditional setting, edges cannot be directed upwards from $\mathbf{X}$ variables to $\mathbf{Y}$, and typically the distribution over $\mathbf{X}$ is not modeled at all. In the causally conditioned setting, edges cannot be directed leftward from future time step variables, $\mathbf{X}_{t+1:T}, \mathbf{Y}_{t+1:T}$ to previous time step variables, $\mathbf{X}_{1:t}, \mathbf{Y}_{1:t}$.

The subtle, but significant difference between the causally conditioned distribution and the conditional probability distribution, $\mathrm{P}(\mathbf{Y}|\mathbf{X}) = \prod_{t=1}^{T} \mathrm{P}(Y_t|\mathbf{X}_{1:T}, \mathbf{Y}_{1:t-1})$—where each Y_t variable is conditioned on all $\mathbf{X}$ variables—is illustrated by Figure 4.2. In the conditional distribution, many edges are directed against the direction of time from Figure 4.1. This distinction serves as the underlying basis for the maximum causal entropy approach.

4.2.3 Causal Entropy and Information

Information-theoretic measures based on the standard conditional probability distribution also extend to the causally conditioned probability distribution (Equation 4.12).

Definition 4.19. *The **causal entropy** (Kramer, 1998; Permuter et al., 2008) of **Y** given **X** is:*

$$\begin{aligned} H(\boldsymbol{Y}^T||\boldsymbol{X}^T) &\triangleq \mathrm{E}_{\boldsymbol{Y},\boldsymbol{X}}[-\log \mathrm{P}(\boldsymbol{Y}^T||\boldsymbol{X}^T)] \\ &= \sum_{t=1}^{T} H(Y_t|\boldsymbol{X}_{1:t}, \boldsymbol{Y}_{1:t-1}). \end{aligned} \tag{4.13}$$

The **causal entropy** (Definition 4.19) measures the uncertainty present in the causally conditioned distribution of the $\mathbf{Y}$ variable sequence given the preceding partial $\mathbf{X}$ variable sequence. It can be interpreted as the expected number of *bits* needed to encode the sequence $\mathbf{Y}_{1:t}$ given the previous $\mathbf{Y}$ variables and sequentially revealed side information, $\mathbf{X}_{1:t}$, which has been revealed at each point in time and excluding unrevealed future side information, $\mathbf{X}_{t+1:T}$.

The **causal information** (also known as the **directed information**) is a measure of the shared information between sequences of variables when the variables are revealed sequentially.

Definition 4.20. *The **causal information** (or **directed information**) of two vectors of variables is:*

$$I(\boldsymbol{X}^T \to \boldsymbol{Y}^T) = \sum_t I(\boldsymbol{X}_{1:t}; Y_t | \boldsymbol{Y}_{1:t-1})$$

It differs from the chain rule for mutual information (Remark 4.12) in that the $\mathbf{X}$ variable vector is limited to t rather than T on the right-hand side.

4.2.4 Properties

Using the causally conditioned probability distributions, any joint distribution can be expressed via the chain rule[3] as $\mathrm{P}(\mathbf{Y}, \mathbf{X}) = \mathrm{P}(\mathbf{Y}^T || \mathbf{X}^T)\ \mathrm{P}(\mathbf{X}^T || \mathbf{Y}^{T-1})$. Our approach estimates $\mathrm{P}(\mathbf{Y}^T || \mathbf{X}^T)$ based on a provided (explicitly or implicitly) distribution of the side information, $\mathrm{P}(\mathbf{X}^T || \mathbf{Y}^{T-1}) = \prod_t \mathrm{P}(X_t | \mathbf{X}_{1:t-1}, \mathbf{Y}_{1:t-1})$. By employing this relationship between the joint and causal distributions, Equation 4.13 can be equivalently expressed in terms of the joint-relative entropy.

Remark 4.21. *The causally conditioned entropy is related to the joint and relative entropy as:*

$$\begin{aligned} H(\boldsymbol{Y}^T || \boldsymbol{X}^T) &= H(\boldsymbol{Y}, \boldsymbol{X}) - H(\boldsymbol{X}^T || \boldsymbol{Y}^{T-1}) \\ &= \textit{const.} - H\left(\boldsymbol{Y}, \boldsymbol{X} // \mathrm{P}_0(\boldsymbol{X}, \boldsymbol{Y})\right) \end{aligned} \tag{4.14}$$

Equation 4.14 is the joint distribution over $\mathbf{X}$ and $\mathbf{Y}$ relative to a baseline joint distribution, $\mathrm{P}_0(\mathbf{X}, \mathbf{Y})$, that obeys the provided side information distribution, $\mathrm{P}(\mathbf{X}^T || \mathbf{Y}^{T-1})$, and is otherwise uniform over $\mathrm{P}(\mathbf{Y}^T || \mathbf{X}^T)$. We employ the "$//$" notation introduced in Definition 4.5 to represent the relative entropy and distinguish it from our employed causal entropy notation, "$||$."

It is easy to verify that the causal entropy upper bounds the conditional entropy; intuitively this reflects the fact that conditioning on information from the future (*i.e.*, acausally) can only decrease uncertainty.

Theorem 4.22. *Generally, the conditional and causal entropies are related as:*

$$H(\boldsymbol{Y}|\boldsymbol{X}) \leq H(\boldsymbol{Y}^T || \boldsymbol{X}^T) \leq H(\boldsymbol{Y}) \leq H(\boldsymbol{Y}, \boldsymbol{X}), \tag{4.15}$$

[3] More generally, other decompositions of the form: $\mathrm{P}(\mathbf{Y}, \mathbf{X}) = \mathrm{P}(\mathbf{Y}^T || \mathbf{X}^{T-\Delta_1})\ \mathrm{P}(\mathbf{X}^T || \mathbf{Y}^{T-\Delta_2})$ for $0 \leq \Delta_1 < \Delta_2$ are possible, including the "standard" conditional decomposition, $P(\mathbf{Y}, \mathbf{X}) = P(\mathbf{X})P(\mathbf{Y}|\mathbf{X})$.

since additional conditioning can never increase entropy. In the special case that $\mathrm{P}(\boldsymbol{X}^T||\boldsymbol{Y}^{T-1})$ *is a deterministic function, then:*

$$H(\boldsymbol{Y}^T||\boldsymbol{X}^T) = H(\boldsymbol{Y}) = H(\boldsymbol{Y}, \boldsymbol{X}). \tag{4.16}$$

Additionally, the causal information obeys inequalities relating it to the mutual information.

Theorem 4.23 (Massey (1990))**.** *The causal information is upper bounded by the mutual information:*

$$I(\boldsymbol{X}_{1:T} \to \boldsymbol{Y}_{1:T}) \leq I(\boldsymbol{X}_{1:T}; \boldsymbol{Y}_{1:T})$$

Theorem 4.24 (Massey (1990))**.** *The causal information is upper bounded by sum of per-timestep mutual information:*

$$I(\boldsymbol{X}_{1:T} \to \boldsymbol{Y}_{1:T}) \leq \sum_{t=1}^{T} I(X_t; Y_t),$$

with equality if and only if $Y_1, ..., Y_T$ *are statistically independent.*

4.2.5 Previous Applications of Causal Information

Causal information theory has found applicability as an information-theoretic measure in the analysis of communication channels with feedback (Massey, 1990; Kramer, 1998), decentralized control (Tatikonda, 2000), and sequential investment and online compression with side information (Permuter et al., 2008). We review those previous results in this section.

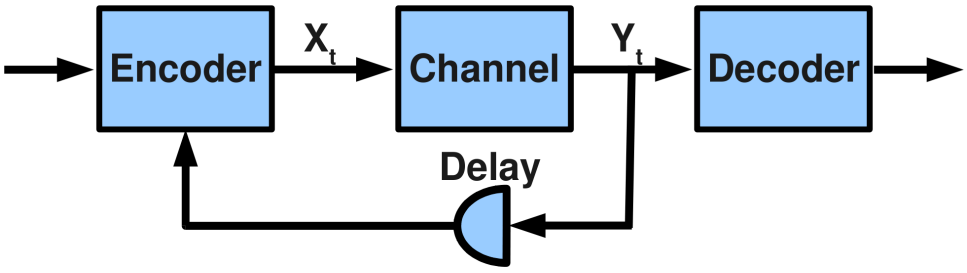

Figure 4.3: A single-user communication channel with delayed feedback.

One of the main focuses of communications research is establishing the maximum amount of information that can be reliably transmitted through a communication channel. Analysis of **channel capacity** (as this quantity is otherwise known) for channels with feedback began with Shannon's result for single-user channels (Figure 4.3) that feedback does not increase channel capacity (Shannon, 1956).

However, in multi-user channels (Shannon, 1961), such as the two-way channel of Figure 4.4, feedback has been shown to increase capacity (Gaarder & Wolf, 1975).

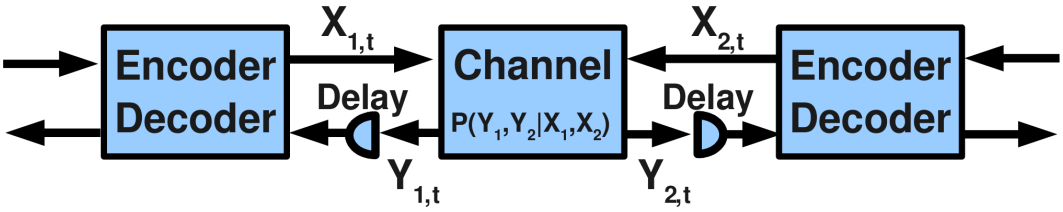

Figure 4.4: A two-way communication channel with delayed feedback.

Theorem 4.25 (Massey (1990))**.** *If $\boldsymbol{X}_{1:T}$ and $\boldsymbol{Y}_{1:T}$ are input and output sequences of a discrete channel, and $\boldsymbol{U}_{1:K}$ is a source output sequence (i.e., message to transmit), then:*

$$I(\boldsymbol{U}_{1:K}; \boldsymbol{Y}_{1:T}) \leq I(\boldsymbol{X}_{1:T} \to \boldsymbol{Y}_{1:T})$$

Combining Theorem 4.23 with Theorem 4.25, Massey (1990) establish the causal information as a tighter bound on the mutual information between the source message and the discrete channel output in communication channels with feedback.

Theorem 4.26 (Corollary 4.1 of Kramer (1998))**.** *The capacity region $\mathcal{C}_{TWC}$ of the discrete memoryless common-output two-way channel is the closure of the set of rate-pairs (R_1, R_2) such that:*

$$R_1 = I(\boldsymbol{X}_{1,1:T} \to \boldsymbol{Y}_{1:T} || \boldsymbol{X}_{2,1:T})$$
$$R_2 = I(\boldsymbol{X}_{2,1:T} \to \boldsymbol{Y}_{1:T} || \boldsymbol{X}_{1,1:T})$$

where T is a positive integer and where:

$$\mathrm{P}(X_{1,t}, X_{2,t} | \boldsymbol{X}_{1,1:t-1}, \boldsymbol{X}_{2,1:t-1}, \boldsymbol{Y}_{1:t-1}) = \mathrm{P}(X_{1,t} | \boldsymbol{X}_{1,1:t-1}, \boldsymbol{Y}_{1:t-1}) \, \mathrm{P}(X_{2,t} | \boldsymbol{X}_{2,1:t-1}, \boldsymbol{Y}_{1:t-1})$$

for all $t = 1, 2, ..., T$.

With Theorem 4.26, Kramer (1998) proves the relationship between directed information and channel capacity in the memory-less common-output two-way channel setting.

Tatikonda (2000) formulates the problem of stochastic control in distributed settings with communication limitations as a problem of selecting the channel input distribution that maximizes the causal information of the channel output $\mathbf{B}_{1:T}$, channel input, $\mathbf{A}_{1:T}$, and the stochastically-controlled state, $\mathbf{Z}_{1:T}$:

$$I(\mathbf{A}_{1:T} \to (\mathbf{B}_{1:T}, \mathbf{Z}_{1:T})).$$

Under a Markovian assumption, a dynamic program that resembles the value iteration (Bellman, 1957) is employed to obtain the channel input distribution.

Permuter et al. (2008) investigate betting strategies in settings where side information providing information on the outcome of horse races is sequentially provided at each point in time (*i.e.*, before each consecutive horse race). They establish the causally conditioned probability distribution as the distribution by which to allocate bets on each of the races to maximize growth rate.

Theorem 4.27 (Permuter et al. (2008))**.** *Given a sequence of horse races with outcomes Y_i at time i with side information X_i at time i, where $o(Y_i|Y^{i-1})$ is the payoff at time i for horse Y_i given that previous horse race winners were $\mathbf{Y}_{1:i-1}$, then for any finite horizon T, the maximum growth rate is achieved when a gambler invests money in proportion to the causal conditioning distribution, i.e.,*

$$b^*(y_i|\mathbf{y}_{1:i-1}, x_i) = p(y_i|\mathbf{y}_{1:i-1}, x^i), \forall \mathbf{y}_{1:i}, x_i, i \leq T. \tag{4.17}$$

The corresponding optimal growth rate is:

$$W^*(\mathbf{Y}^T||\mathbf{X}^T) = E[\log o(Y^{1:T})] - H(\mathbf{Y}^T||\mathbf{X}^T). \tag{4.18}$$

The result of Theorem 4.27 extends maximum growth rate investing (Theorem 4.16 and Theorem 4.17) to the sequential setting with variable dependence.

4.3 Discussion

In this chapter, we reviewed the well-known information-theoretic measures of entropy and information for marginal, joint, conditional, and relative probability distributions. We then reviewed causal information theory, an extension of information theory to settings with interaction and feedback. We summarized the key results of the applications of causal information theory for communication, control, and gambling problems. In Chapter 5, we connect the standard information-theoretic concepts to existing techniques for prediction. We then extend those techniques using the less widely known causal information theory for prediction of purposeful, adaptive behavior.

Chapter 5

The Principle of Maximum Causal Entropy

"It is far better to foresee even without certainty than not to foresee at all."
— Henri Poincaré (Mathematician, 1854–1912).

In this chapter, we motivate and present **the principle of maximum causal entropy** (Ziebart et al., 2010b). This principle extends the maximum entropy approach (Jaynes, 1957) to conditional probability distributions in settings characterized by *interaction with stochastic processes where side information from those processes is dynamic*, *i.e.*, revealed over time according to a known conditional probability distribution. Importantly, each future side information variable is latent during earlier points of interaction, and its particular instantiation should have no causal influence before it is revealed. Instead, the principle of maximum causal entropy prescribes expectations over future side information variables while they are latent. Of significance for the focus of this thesis, this framework provides a principled approach with predictive guarantees for modeling observed behavior sequences characterized by features or statistics that define its purposefulness or rationality, while being sequential influenced by external, random factors.

We begin this chapter by reviewing the principle of maximum entropy. We summarize some of its important properties and establish its connection to probabilistic graphical models. We then employ causal information theory (Chapter 4) to extend the principle of maximum entropy to sequential settings with adaptation, feedback, and partial information.

5.1 Principle of Maximum Entropy

When given only partial information about a probability distribution, $\tilde{P}$, typically many different distributions, P, are capable of matching that information. For example, many distributions have the same mean value. The principle of maximum entropy resolves the ambiguity of an under-constrained distribution by selecting the single distribution that has the least *commitment* to any particular outcome while matching the observational constraints imposed on the distribu-

tion (Jaynes, 1957). The lack of commitment (or uncertainty) of the distribution is measured by Shannon's information entropy (Definition 4.1).

Definition 5.1. *The **maximum entropy probability distribution** is the distribution with maximum information entropy subject to matching expected characteristics (denoted as characteristic functions $g_k(P)$ with empirical characteristic averages $g_k(\tilde{P})$) and is obtained from the following optimization:*

$$\begin{aligned} &\underset{\mathrm{P}}{\operatorname{argmax}}\, H_{\mathrm{P}}(X) \\ \textit{such that: } &\forall_k\ g_k(P) = g_k(\tilde{P}) \\ &\sum_x \mathrm{P}(x) = 1 \\ &\forall_x\ \mathrm{P}(x) \geq 0 \end{aligned} \tag{5.1}$$

While any characteristic function is possible, we will focus on probabilistically weighted characteristic functions, $g_k(P) = \mathrm{E}_X[f_k(x)] = \sum_{x \in \mathcal{X}} \mathrm{P}(x) f_k(x)$. The distribution satisfying the maximization then has the following form:

$$\mathrm{P}(x) = Z(\alpha)^{-1}\ e^{\sum_k \alpha_k f_k(x)}. \tag{5.2}$$

The parameters, $\{\alpha_k\}$, are generally obtained using numerical optimization techniques to satisfy the constraints of Equation 5.1. The normalization factor, $Z(\alpha) = \sum_x e^{\sum_k \alpha_k f_k(x)}$, forces the distribution to normalize and is also called the **partition function**.

Many of the fundamental building blocks of statistics are maximum entropy distributions. The Gaussian distribution, for example, has the form of a maximum entropy distribution constrained to match first and second moments. Similarly, more sophisticated probabilistic models, such as Markov random fields, also maximize the entropy subject to similar constraints, as we explore in Section 5.1.3.

5.1.1 Justifications for the Principle of Maximum Entropy

There are a number of different arguments that have been put forth to justify the principle of maximum entropy's usage. Jaynes (1957) originally employed an information-theoretic justification that a lower entropy probability distribution corresponds to "committing" to enforce additional assumptions on the probability distribution that are not necessary given the available information (*i.e.*, constraints on the distribution). While this justification has an intuitive appeal as a generalization of the principle of indifference, other arguments present stronger quantitative justification.

A second justification, suggested by Graham Wallis to Edwin Jaynes, derives the maximum entropy distribution as the most probable of all "fair" random distributions that match provided constraints (Jaynes & Bretthorst, 2003, Chapter 11). Consider the process of assigning probabilities $\{p_1, ..., p_m\}$ to m different possibilities by (uniformly) randomly assigning $n \gg m$ quanta

among those possibilities. If the random assignment does not match provided constraints on the distribution, a new assignment is generated until one that matches constraints is found. The most likely distribution that will be found by this process maximizes $W = \frac{n!}{n_1!\ldots n_m!}$. As $n \to \infty$, a monotonic function of W, $\frac{1}{n}\log W \to -\sum_{i=1}^{m} p_i \log p_i$, which is the entropy of the distribution. As a result, the most likely constraint-satisfying assignment from this process is equivalent to the maximum entropy distribution in the limit of infinitesimally small quanta of probability. Note that this derivation neither assumes nor employs any notions from information theory. Rather, information-theoretic measures follow from the derivation.

An important predictive justification is from a game-theoretic perspective. It pits the decision maker and nature against each other in the choice of probability distributions $\mathrm{P}(X)$ and $\tilde{P}(X)$, as shown in Theorem 5.2.

Theorem 5.2 (Grünwald & Dawid (2003))**.** *The **maximum entropy distribution** minimizes the worst case prediction log-loss,*

$$\inf_{\mathrm{P}(X)} \sup_{\tilde{\mathrm{P}}(X)} -\sum_{X} \tilde{\mathrm{P}}(X) \log \mathrm{P}(X),$$

given e.g., feature expectation constraints that P and $\tilde{P}$ both match: $\mathrm{E}_{\tilde{\mathrm{P}}(X)}[\mathcal{F}(X)]$.

This can be viewed as a two-step game where the decision maker first selects a probability distribution, P, and then nature adversarially chooses the empirical distribution, $\tilde{P}$. Both distributions are constrained to match provided constraints. The maximum entropy distribution is the one where the decision maker chooses P to minimize the worst log-loss that nature can possible create by choosing $\tilde{P}$ adversarially. As the log prediction rate is a natural evaluation metric for machine learning applications, minimizing its worst case value is an important guarantee for machine learning techniques. The extension of this guarantee to sequential settings representing behavior is one of the central contributions of this thesis.

Lastly, there is a useful gambling justification for maximum entropy. Returning to the horse racing example of Section 4.1.3, the doubling rate is defined as:

$$\begin{aligned} W(\mathbf{X}) &= \sum_{x} \mathrm{P}(x) \log \left(b(x)\ o(x)\right) \\ &= \sum_{x} \left(\mathrm{P}(x) \log b(x) + \mathrm{P}(x) \log o(x)\right). \end{aligned} \tag{5.3}$$

When the payoffs, $o(\mathbf{X})$, are uniform, the right-hand expression of Equation 5.3 reduces to a constant that is independent of the bet distribution, b. We consider the case where some properties that $\mathrm{P}(\mathbf{X})$ must satisfy, *i.e.*, $E_{\tilde{\mathrm{P}}}[\mathcal{F}(Y)]$, are known. Following the same adversarial viewpoint of Theorem 5.2, the maximum entropy distribution for $b(\mathbf{X})$ maximizes the worst-case doubling rate,

$$b(\mathbf{X})^* = \operatorname*{argmax}_{b(\mathbf{X})} \min_{\mathrm{P}(\mathbf{X})} W(\mathbf{X}),$$

where $\mathrm{P}(\mathbf{X})$ is chosen adversarially by "nature," but must match $E_{\tilde{\mathrm{P}}}[\mathcal{F}(Y)]$.

5.1.2 Generalizations of the Principle of Maximum Entropy

The principle of maximum entropy can be generalized by using alternate entropy measures. For example, the conditional entropy (Equation 4.5) or relative entropy (KL divergence, Equation 4.6) can be optimized in lieu of the standard entropy. We shall make use of and extend the conditional entropy generalization throughout this thesis. On the other hand, we argue that optimizing the relative entropy provides no additional generality in the maximum entropy framework, and that it need not be considered as we generalize the entropy to settings with feedback throughout the remainder of this thesis.

A choice of relative distribution, $Q(x)$, is required to optimize the relative entropy. The distribution obtained by optimizing the relative entropy then takes the following form:

$$\begin{aligned} \mathrm{P}(x) &\propto Q(x) e^{\sum_k \alpha_k f_k(x)} \\ &= e^{\log Q(x) + \sum_k \alpha_k f(x)} \end{aligned} \tag{5.4}$$

Note that Equation 5.4 could be similarly realized by introducing a new feature, $f_Q(x) = \log Q(x)$, into the maximum entropy probability distribution (Equation 5.2) and assigning it a fixed weight of $\alpha_Q = 1$. If the relative distribution is given, *e.g.*, by physical laws, this approach can be directly applied. However, we argue that an appropriate relative distribution, $Q(x)$, is generally not known. Instead, a set of candidate relative distributions, $\{Q_1(x), Q_2(x), ...\}$, may be available and an appropriate set of weights $(\alpha_{Q_1}, \alpha_{Q_2}, ...)$ should be learned for those candidate relative distributions rather than assumed. We more formally define this notion in Remark 5.3.

Remark 5.3. *The maximum entropy distribution for* $\mathrm{P}(x)$ *(Equation 5.2) augmented with* $\log Q(x)$ *features (corresponding to constraints on* $\mathrm{E}_{\mathrm{P}}[\log Q(x)]$*) is at least as expressive as the distribution optimizing the entropy of* $\mathrm{P}(x)$ *relative to* $Q(x)$*. Equivalence is realized when the parameter* α_Q *associated with the augmented feature in the maximum entropy model is fixed to 1.0 and the remaining parameters are obtained through maximum entropy optimization. The maximum entropy distribution can be more general when the choice of weights* $\{\alpha_Q\}$ *is also optimized.*

Since the relative entropy affords no greater generality than this simple extension of the standard maximum entropy approach, we will not consider it as a meaningful generalization as we present the maximum causal entropy approach.

5.1.3 Probabilistic Graphical Models as Entropy Maximization

Existing probabilistic graphical models can be formulated from the maximum entropy approach. This perspective is important for understanding the generalization that the maximum causal entropy approach of this thesis provides. Markov random fields can be interpreted as an extension of maximum entropy to joint probability distributions.

The form of the Markov random field (Definition 2.3) can be derived as a maximum entropy optimization (with probabilistic constraints on P(**X**) suppressed) as follows:

$$\max_{\mathrm{P}(\mathbf{X})} H(\mathbf{X})$$
$$\text{such that: } \forall_i \ \mathrm{E}_{\mathrm{P}(\mathbf{X})}[f_{C_i}(\mathbf{X}_{C_i})] = \mathrm{E}_{\tilde{\mathrm{P}}(\mathbf{X})}[f_{C_i}(\mathbf{X}_{C_i})],$$

where $\tilde{\mathrm{P}}$ represents empirical probability distribution variables obtained from observed data, and $\{C_i\}$ are sets of subsets of $\mathbf{X}$ (*i.e.*, cliques in the corresponding undirected graphical representation). The maximum entropy log-loss guarantee extends to the Markov random field.

Corollary 5.4 (of Theorem 5.2)**.** *Markov random fields guarantee minimal worst-case predictive log-loss of $\boldsymbol{X}_{1:T}$ subject to constraints $\{f_{C_i}\}$.*

Conditional random fields are derived from the maximum entropy approach by maximizing the conditional entropy (Definition 4.4), $H(\mathbf{Y}|\mathbf{X})$, subject to feature function expectation constraints (with feature functions over $\mathbf{X}$ and $\mathbf{Y}$):

$$\max_{\mathrm{P}(\mathbf{Y}|\mathbf{X})} H(\mathbf{Y}|\mathbf{X})$$
$$\text{such that: } \forall_i \ \mathrm{E}_{\mathrm{P}(\mathbf{Y},\mathbf{X})}[f_{C_i}(\mathbf{Y}_{C_i}, \mathbf{X}_{C_i})] = \mathrm{E}_{\tilde{\mathrm{P}}(\mathbf{Y},\mathbf{X})}[f_{C_i}(\mathbf{Y}_{C_i}, \mathbf{X}_{C_i})],$$

where $P(\mathbf{Y}, \mathbf{X}) = P(\mathbf{Y}|\mathbf{X})\tilde{P}(\mathbf{X})$. The conditional random field provides a predictive log-loss minimization guarantee for its information availability setting.

Corollary 5.5 (of Theorem 5.2)**.** *Conditional random fields guarantee minimal worst-case predictive log-loss of $\boldsymbol{Y}_{1:T}$ when given side information, $\boldsymbol{X}_{1:T}$, up front and constraints, $\{f_{C_i}\}$, governing the probability distribution.*

The maximum causal entropy approach of this thesis generalizes this guarantee to settings where side information, $\mathbf{X}_{1:T}$, is revealed sequentially.

5.1.4 Approximate Constraints and Bayesian Inference

A number of criticisms of the principle of maximum entropy illustrate incompatibilities between employing the principle of maximum entropy and Bayesian inference. Jaynes distinguishes the two by stating that the principle of maximum entropy "assigns" a distribution while applying Bayes' theorem means "calculating" a probability (Jaynes, 1988). Formally reconciling or explaining those differences is beyond the scope of this thesis. Instead we offer two high-level considerations to alleviate some of the concerns over these incompatibilities. In the remainder of this thesis, we will employ both the maximum entropy principle and Bayesian inference techniques as appropriate based on the modeling or inference task.

The first consideration is that the constraints of maximum entropy optimization are often assumed to be absolute. Instead, in practice they are often estimates from finite samples of data. Dudík & Schapire (2006) show that appropriately incorporating approximate constraints corresponds to regularization function, $R(\alpha)$, in the dual optimization objective:

$$\underset{\lambda}{\operatorname{argmax}} \left(\sum_t \log \mathrm{P}(x_t|\lambda) \right) - R(\lambda). \tag{5.5}$$

The regularization term can be generally interpreted as a prior on the probability distribution. Its form can be made to match common probability distributions used as priors in Bayesian inference approaches depending on the noise assumptions of the constraints on the primal optimization. Two such primal approximations on the estimates and corresponding supplementary dual potentials are shown in Table 5.1.

Table 5.1: Primal approximation potentials and dual regularization terms from Dudík & Schapire (2006).

Description	**Primal constraint/potential**	**Dual potential**
Box constraints	$\forall_j \lvert \mathrm{E}_{\tilde{\mathrm{P}}}[f_j] - \mathrm{E}_{\mathrm{P}}[f_j] \rvert \leq \beta_j$	$\lambda^\top \mathrm{E}_{\tilde{\mathrm{P}}}[\mathbf{f}] + \sum_j \beta_j \lvert \lambda_j \rvert$
l_2^2 norm	$\frac{\lvert\lvert \mathrm{E}_{\tilde{\mathrm{P}}}[\mathbf{f}] - \mathrm{E}_{\mathrm{P}}[\mathbf{f}] \rvert\rvert_2^2}{(2\alpha)}$	$\lambda^\top \mathrm{E}_{\tilde{\mathrm{P}}}[\mathbf{f}] + \alpha \lvert\lvert \lambda \rvert\rvert_2^2$

The second consideration is the difference between applying new information in parallel and applying it in sequence. For example, consider a six-sided die with expected value, $\mathrm{E}[X] = 3.5$. If new information is provided that *e.g.*, a given dice roll is even, $X \equiv 0 \pmod 2$, an undesirable result is obtained by finding the maximum entropy distribution that satisfies both constraints—P(X=2) is more probable than P(X=6) in that case. Instead, approaches to reconcile maximum entropy and Bayesian inference suggest first obtaining a distribution satisfying the initial constraint and then maximizing the entropy of a new distribution constrained to match the added constraint relative to that first distribution (Giffin & Caticha, 2007). In other words, constraints should be applied in sequence.

5.2 Maximum Causal Entropy

In this section, we employ causal information-theoretic measures to extend the maximum entropy approach to settings with sequential information revelation. These include partial information and partial controllability settings. This extension enables extension of the maximum entropy approach to purposeful, adaptive behavior settings investigated in this thesis.

5.2.1 Convex Optimization

With the causal entropy (Equation 4.13) as our objective function, we now pose and analyze the maximum causal entropy optimization problem.

Definition 5.6. *The **general form** of the **principle of maximum causal entropy** prescribes the causally conditioned entropy-maximizing probability distribution,* $\mathrm{P}(\boldsymbol{Y}^T||\boldsymbol{X}^T) \triangleq \{\mathrm{P}(\boldsymbol{Y}^T||\boldsymbol{X}^T)\}$ *consistent with a set of affine equality constraints,* $\{g_i(\mathrm{P}(\boldsymbol{Y}^T||\boldsymbol{X}^T)) = 0\}$*, and convex inequality constraints,* $\{h_j(\mathrm{P}(\boldsymbol{Y}^T||\boldsymbol{X}^T)) \leq 0\}$*, and a given distribution of side information,* $\mathrm{P}(\boldsymbol{X}^T||\boldsymbol{Y}^{T-1})$*. Obtaining* $\mathrm{P}(\boldsymbol{Y}^T||\boldsymbol{X}^T)$ *can be accomplished by the following optimization:*

$$
\begin{aligned}
&\underset{\mathrm{P}(\boldsymbol{Y}^T||\boldsymbol{X}^T)}{\operatorname{argmax}} \; H(\boldsymbol{Y}^T||\boldsymbol{X}^T) \\
\textit{such that: } &\forall_i \; g_i\left(\mathrm{P}(\boldsymbol{Y}^T||\boldsymbol{X}^T)\right) = 0, \\
&\forall_j \; h_j\left(\mathrm{P}(\boldsymbol{Y}^T||\boldsymbol{X}^T)\right) \leq 0, \\
&\forall_{\boldsymbol{X},\boldsymbol{Y}} \; \mathrm{P}(\boldsymbol{Y}^T||\boldsymbol{X}^T) \geq 0, \\
&\forall_{\boldsymbol{X}} \sum_{\boldsymbol{Y}} \mathrm{P}(\boldsymbol{Y}^T||\boldsymbol{X}^T) - 1 = 0, \textit{ and} \\
&\forall_{\tau,\boldsymbol{Y}_{1:\tau},\boldsymbol{X},\boldsymbol{X}':\boldsymbol{X}_{1:\tau}=\boldsymbol{X}'_{1:\tau}} \sum_{\boldsymbol{Y}_{\tau+1:T}} \left(\mathrm{P}(\boldsymbol{Y}^T||\boldsymbol{X}^T) - \mathrm{P}(\boldsymbol{Y}^T||\boldsymbol{X}'^T)\right) = 0,
\end{aligned}
\tag{5.6}
$$

where $P(\mathbf{Y}^T||\mathbf{X}^T)$ represents the set of random variables: $\{P(\mathbf{y}_{1:T}||\mathbf{x}_{1:T})\}$.

We refer to the first two constraints of the general maximum causal entropy optimization (Equation 5.6)—the sets of affine equality constraints (g_i) and convex inequality constraints (h_i)—as **behavioral constraints**. These constraints vary between problem settings and domains, but typically the equality constraints can be interpreted as providing *efficiency guarantees*, while the inequality constraints provide *rationality guarantees*. The remaining constraints are the same for all maximum causal entropy problem settings: the third and fourth are **probabilistic constraints** that enforce non-negativity and normalization, and the final constraints are **causal constraints** that prevent the influence of future side information on previous $\mathbf{Y}$ variables. More specifically, this final causal constraint forces the causally conditioned probability up to point τ in time to be equivalent for all possible future side information values of $\mathbf{X}_{\tau+1:T}$.

Remark 5.7. *The causal constraints of Equation 5.6 force the causally conditioned probability to factor as* $\mathrm{P}(\boldsymbol{Y}^T||\boldsymbol{X}^T) = \prod_t \mathrm{P}(Y_t|\boldsymbol{Y}_{1:t-1}, \boldsymbol{X}_{1:t})$*; otherwise, the conditional probability,* $P(\boldsymbol{Y}|\boldsymbol{X})$*, would result. While future* $\boldsymbol{X}$ *variables and past* $\boldsymbol{Y}$ *variables are not necessarily statistically independent, future* $\boldsymbol{X}$ *variables do not causally influence past* $\boldsymbol{Y}$ *variables due to this forced factorization.*

Theorem 5.8. *The general maximum causal entropy optimization (Equation 5.6)—or, more correctly, minimizing its negation—is a convex optimization.*

Proof (sketch). The negative objective is a linear combination of convex functions of the form $Cx \log x$. The behavioral constraints by specification define a convex constraint space and the remaining constraints enforcing causality requirements are affine in terms of $P(\mathbf{Y}^T||\mathbf{X}^T)$ variables. □

Though standard convex optimization techniques (Boyd & Vandenberghe, 2004) can be employed to solve this general maximum causal entropy primal optimization, when T, $|\mathcal{Y}|$, and/or $|\mathcal{X}|$ are large, the number of constraints required to prevent future latent variables from influencing earlier variables can grow exponentially. Specialized optimization techniques for specific forms of $\{g_i(\mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T))\}$ and $\{h_j(\mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T))\}$ can provide significant computational efficiency improvements by decomposing the problem into a sequence of easier sub-optimizations, as we shall see in later chapters and applications in this thesis.

5.2.2 Convex Duality

Often optimizing the dual of the maximum causal entropy optimization (Equation 5.6) leads to desired computational efficiency improvements. This is particularly true when the number of behavioral constraints is significantly smaller than the causal conditional probability variables, $\{\mathrm{P}(\mathbf{Y}||\mathbf{X})\}$. The general form of the dual is as follows:

$$\inf_{\mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T)} -H(\mathbf{Y}^T||\mathbf{X}^T) + \sum_{i=1}^{m} \lambda_i g_i(\mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T)) + \sum_{j=1}^{n} \gamma_j h_j(\mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T)) \tag{5.7}$$

$$+ \sum_{\mathbf{X}} C_{\mathbf{X}} \left(\sum_{\mathbf{Y}} \mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T) - 1 \right) + \sum_{\tau, \mathbf{Y}_{1:\tau}, \mathbf{X}, \mathbf{X}' : \mathbf{X}_{1:\tau} = \mathbf{X}'_{1:\tau}} \phi_{\tau, \mathbf{Y}, \mathbf{X}, \mathbf{X}'} \left(\sum_{\mathbf{Y}^{\tau+1:T}} \mathrm{P}(\mathbf{Y}||\mathbf{X}) - \mathrm{P}(\mathbf{Y}^T||\mathbf{X}'^T) \right),$$

where dual parameters are unbounded λ and $\gamma \geq 0$.

The final set of constraints force the causal conditional distribution to factor according to $\mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T) = \prod_t \mathrm{P}(Y_t|\mathbf{Y}_{1:t-1}, \mathbf{X}_{1:t})$ as described by Remark 5.7. It is often convenient to enforce those constraints by maximizing the dual in terms of those $\mathrm{P}(Y_t|\mathbf{Y}_{1:t-1}, \mathbf{X}_{1:t})$ factors:

$$\inf_{\{\mathrm{P}(Y_t|\mathbf{Y}_{1:t-1}, \mathbf{X}_{1:t})\}} - H(\mathbf{Y}^T||\mathbf{X}^T) + \sum_{i=1}^{m} \lambda_i g_i(\mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T)) + \sum_{j=1}^{n} \gamma_j h_j(\mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T)), \tag{5.8}$$

with Lagrangian parameters λ and $\gamma \geq 0$. Here we have also dropped the probabilistic normalization constraint for brevity with the implicit understanding that the resulting causal conditional distribution must normalize and that each $\mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T)$ term and each $\mathrm{P}(Y_t|\mathbf{Y}_{1:t-1}, \mathbf{X}_{1:t})$ term must have non-negative value.

Typically, a parametric form for $P(Y_t|\mathbf{Y}_{1:t-1}, \mathbf{X}_{1:t})$ can be obtained by solving for the closed-form probability in terms of parameters λ and γ.

Theorem 5.9. *Strong duality holds for the maximum causal entropy optimization.*

The technical considerations for duality are discussed in the proof of Theorem 5.9 in Section A.2.

5.2.3 Worst-Case Predictive Guarantees

The probability distribution that maximizes the causal entropy provides important worst-case guarantees when predicting the future $\mathbf{Y}$ variables given only previously available $\mathbf{X}$ and $\mathbf{Y}$ variables and constraints on the probability distribution.

Theorem 5.10. *The maximum causal entropy distribution minimizes the worst-case prediction log-loss,*

$$\inf_{\mathrm{P}(\boldsymbol{Y}||\boldsymbol{X})} \sup_{\tilde{\mathrm{P}}(\boldsymbol{Y}^T||\boldsymbol{X}^T)} -\sum_{\boldsymbol{Y},\boldsymbol{X}} \tilde{\mathrm{P}}(\boldsymbol{Y},\boldsymbol{X}) \log \mathrm{P}(\boldsymbol{Y}^T||\boldsymbol{X}^T), \tag{5.9}$$

given that $\tilde{\mathrm{P}}(\boldsymbol{Y},\boldsymbol{X}) = \tilde{\mathrm{P}}(\boldsymbol{Y}^T||\boldsymbol{X}^T)\,\mathrm{P}(\boldsymbol{X}^T||\boldsymbol{Y}^{T-1})$ *and sets of constraints* $\forall_i \;\; g_i(\mathrm{P}(\boldsymbol{Y}^T||\boldsymbol{X}^T)) = 0$, $\forall_j \;\; h_j(\mathrm{P}(\boldsymbol{Y}^T||\boldsymbol{X}^T)) \leq 0$*, when* $\boldsymbol{X}$ *is sequentially revealed from the known causally conditioned distribution of side information,* $\mathrm{P}(\boldsymbol{X}^T||\boldsymbol{Y}^{T-1})$.

Theorem 5.10 follows naturally from Theorem 5.2 (Grünwald & Dawid, 2003) and extends the "robust Bayes" results to the interactive setting as the main justification for employing the maximum causal entropy approach for predictive applications. The theorem can be understood by viewing maximum causal entropy as a *maximin* game where nature chooses a distribution to maximize a predictor's causal uncertainty while the predictor tries to minimize it. By duality, the *minimax* view of the theorem is equivalent.

The view in this thesis is that the constraints of the maximum causal entropy optimization should capture the structured, purposeful qualities of observed behavior. In that case, the worst-case predictive log-loss guarantees of Theorem 5.10 force the distribution to be as uncertain as possible about "non-purposeful" qualities of behavior. If those aspects of behavior are not purposeful, the agent should be indifferent to possible options that differ only in terms of those aspects.

5.2.4 Gambling Growth Rate Guarantees

The principle of maximum causal entropy also extends the growth rate guarantees for gambling to the interactive setting. Consider side information variables that are sequentially revealed in a series of horse races. Further, assume that the side information that is revealed depends on the previous horse races (and bets).

The doubling rate for this causal setting is:

$$\begin{aligned} W(\mathbf{Y}||\mathbf{X}) &= \sum_{\mathbf{y},\mathbf{x}} P(\mathbf{y},\mathbf{x}) \log \left(b(\mathbf{y}||\mathbf{x}) o(\mathbf{y}||\mathbf{x}) \right) \\ &= \sum_{\mathbf{y},\mathbf{x}} \left(P(\mathbf{y},\mathbf{x}) \log b(\mathbf{y}||\mathbf{x}) + P(\mathbf{y},\mathbf{x}) \log o(\mathbf{y}||\mathbf{x}) \right) \end{aligned} \tag{5.10}$$

We can ignore the right-hand term of Equation 5.10, which is again a constant for appropriate "uniform" choice of odds. We assume equality and inequality constraints, $\{g_i\}$ and $\{h_j\}$, on

the probability distribution, $\mathrm{P}(\mathbf{Y}, \mathbf{X})$. Additionally, the distribution must obey the known side information dynamics, $\mathrm{P}(\mathbf{X}^T||\mathbf{Y}^{T-1})$. Maximizing the worst-case growth rate for bet distribution, $\mathbf{b}$, is then:

$$b(\mathbf{Y}||\mathbf{X})^* = \underset{b(\mathbf{y}||\mathbf{x})}{\operatorname{argmax}} \min_{\mathrm{P}(\mathbf{y},\mathbf{x})} W(\mathbf{Y}||\mathbf{X}). \tag{5.11}$$

Again, $\mathrm{P}(\mathbf{Y}, \mathbf{X})$ is chosen adversarially, but must satisfy the provided equality and inequality constraints. This betting distribution is a maximum causal entropy distribution, which can be noted as a consequence of Theorem 5.10. (The left-hand side of Equation 5.10 is the negative of Equation 5.9.)

5.3 Information-Theoretic Extensions and Special Cases

5.3.1 Static Conditioning Extension

We formulated the maximum causal entropy optimization with causal conditioning in Section 5.2.1. Causal conditioning is distinguished from typical conditioning, which we will refer to as **static conditioning** in this work, by the ordered constraints on the causal influence of side information variables. However, in many problems both causal conditioning and static conditioning are appropriate. The distribution of $\mathbf{Y}$ causally conditioned on $\mathbf{X}$ and statically conditioned on $\mathbf{W}$ is notated and defined (Kramer, 1998) as:

$$\mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T|\mathbf{W}) \triangleq \prod_t \mathrm{P}(Y_t|\mathbf{Y}_{1:t-1}, \mathbf{X}_{1:t}, \mathbf{W}). \tag{5.12}$$

In the context of decisions and control problems, this static information may correspond to structure and/or characteristics of the decision problem that vary between instances. The entropy measure corresponding to the static and causally conditioned entropy is:

$$H(\mathbf{Y}^T||\mathbf{X}^T|\mathbf{W}) \triangleq \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y}|\mathbf{W})}\left[-\log \mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T|\mathbf{W})\right]. \tag{5.13}$$

The previously employed convex equality and inequality behavioral constraints can be conditioned on $\mathbf{W}$, as well as the side information dynamics conditional distribution, $\mathrm{P}(\mathbf{X}^T||\mathbf{Y}^{T-1}|\mathbf{W})$. The optimization problem for this static and causally conditioned distribution is then:

$$\begin{aligned}
&\operatorname*{argmax}_{\mathrm{P}(\mathbf{Y}||\mathbf{X}|\mathbf{W})} H(\mathbf{Y}^T||\mathbf{X}^T|\mathbf{W}) \qquad (5.14)\\
\text{such that: } &\forall_i\ g_i(\mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T|\mathbf{W}), \mathbf{W}) = 0,\\
&\forall_j\ h_j(\mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T|\mathbf{W}), \mathbf{W}) \leq 0,\\
&\forall_{\mathbf{X},\mathbf{Y},\mathbf{W}}\ \mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T|\mathbf{W}) \geq 0,\\
&\forall_{\mathbf{X},\mathbf{W}}\ \sum_{\mathbf{Y}} \mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T|\mathbf{W}) - 1 = 0, \text{ and}\\
&\forall_{\tau,\mathbf{Y}_{1:\tau},\mathbf{W},\mathbf{X},\mathbf{X'}:X_{1:\tau}=X'_{1:\tau}} \sum_{\mathbf{Y}_{\tau+1:T}} \Big(\mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T|\mathbf{W}) - \mathrm{P}(\mathbf{Y}^T||\mathbf{X'}^T|\mathbf{W})\Big) = 0.
\end{aligned}$$

The factored dual (Equation 5.8) naturally extends to this setting with distributions and constraints conditioned on $\mathbf{W}$ (again, with probabilistic normalization multipliers suppressed):

$$\inf_{\{\mathrm{P}(Y_t|\mathbf{Y}_{1:t-1},\mathbf{X}_{1:t},\mathbf{W})\}} -H(\mathbf{Y}^T||\mathbf{X}^T|\mathbf{W}) + \sum_{i=1}^{m} \lambda_i g_i(\mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T|\mathbf{W})) + \sum_{j=1}^{n} \gamma_j h_j(\mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T|\mathbf{W})), \quad (5.15)$$

with $\gamma \geq 0$. Similarly, the worst-case log loss guarantee of Theorem 5.10 also extends to this setting by appropriately conditioning on $\mathbf{W}$. In many settings, we consider $\mathbf{W}$ to be a set of static variables describing the environment or decision setting. However, it is often convenient notationally to suppress this static conditioning.

5.3.2 Optimizing Relative Causal Entropy

The principle of maximum causal entropy can be viewed as optimizing the causal entropy of action sequences relative to a uniform distribution over action sequences. A natural attempt to generalize is to optimize the causal entropy relative to a baseline distribution $\mathrm{P}_0(\mathbf{Y}||\mathbf{X})$. We denote this measure as:

$$\begin{aligned}
H(\mathbf{Y}^T||\mathbf{X}^T // \mathrm{P}_0(\mathbf{Y}^T||\mathbf{X}^T)) &\triangleq \mathrm{E}_{\mathrm{P}(\mathbf{Y},\mathbf{X})}\left[\log \frac{\mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T)}{\mathrm{P}_0(\mathbf{Y}^T||\mathbf{X}^T)}\right]\\
&= \mathrm{E}_{\mathrm{P}(\mathbf{Y},\mathbf{X})}\left[\log \mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T)\right] + \mathrm{E}_{\mathrm{P}(\mathbf{Y},\mathbf{X})}\left[-\log \mathrm{P}_0(\mathbf{Y}^T||\mathbf{X}^T)\right]. \quad (5.16)
\end{aligned}$$

However, by expanding the causal relative entropy (Equation 5.16), we can see that by following the general argument of Remark 5.3, no additional generality is provided by considering the relative causal entropy dual optimization (Remark 5.11).

Remark 5.11. *A maximum causal relative entropy distribution that is relative to a causally conditioned distribution* $\mathrm{P}_0(\boldsymbol{Y}||\boldsymbol{X})$*,*

$$\max_{\lambda,\gamma\geq 0} \inf_{\{\mathrm{P}(Y_t|\boldsymbol{Y}_{1:t-1},\boldsymbol{X}_{1:t})\}} - H(\boldsymbol{Y}^T||\boldsymbol{X}^T) + \mathrm{E}_{\mathrm{P}(\boldsymbol{Y},\boldsymbol{X})}\left[\log \mathrm{P}_0(\boldsymbol{Y}^T||\boldsymbol{X}^T)\right] + \sum_{i=1}^{m} \lambda_i g_i(\mathrm{P}(\boldsymbol{Y}^T||\boldsymbol{X}^T)) + \sum_{j=1}^{n} \gamma_j h_j(\mathrm{P}(\boldsymbol{Y}^T||\boldsymbol{X}^T)), \tag{5.17}$$

can be obtained in the maximum causal entropy model by incorporating an additional affine Lagrange multiplier term, $\lambda_{\mathrm{P}_0}\, \mathrm{E}_{\mathrm{P}(\boldsymbol{Y},\boldsymbol{X})}[\log \mathrm{P}_0(\boldsymbol{Y}^T||\boldsymbol{X}^T)]$*, where the multiplier* λ_{P_0} *is fixed to 1, equating Equation 5.17 and Equation 5.8.*

Any relative distribution can be incorporated into a maximum causal entropy model without explicitly using the relative entropy by augmenting with additional features (Remark 5.11). However, we generally advocate learning the weights of all of the Lagrange multipliers from data rather than assuming their values. This allows us to consider many possible relative distributions and to learn which one or what combination of them is most explanatory of demonstrated data.

5.3.3 Continuous-Valued Maximum Causal Entropy

Many important problems in estimation and control are characterized by actions and states within continuous spaces. In principle, these continuous spaces can be discretized into finer and finer discrete resolutions with smaller and smaller approximation inaccuracies. Then discrete maximum causal entropy methods can be employed. However, reasoning in the continuous space of states and actions can be computationally beneficial for some problems.

We will limit our focus to continuous problem with causally conditioned probability density functions, $P^*_{\lambda,\gamma}(\mathbf{Y}^T||\mathbf{X}^T)$, that are closed-form functions of the Lagrange multipliers λ and γ. The dual optimization problems are then of the form:

$$\max_{\lambda,\gamma\geq 0} \mathrm{E}_{\mathrm{P}(\mathbf{Y},\mathbf{X})}[-\log P^*_{\lambda,\gamma}(\mathbf{Y}^T||\mathbf{X}^T)] + \sum_{i=1}^{m} \lambda_i g_i(P^*_{\lambda,\gamma}(\mathbf{Y}^T||\mathbf{X}^T)) + \sum_{j=1}^{n} \gamma_j h_i(P^*_{\lambda,\gamma}(\mathbf{Y}^T||\mathbf{X}^T)), \tag{5.18}$$

where $\mathrm{E}_{\mathrm{P}(\mathbf{Y},\mathbf{X})}[f(\mathbf{X},\mathbf{Y})] = \int_{\mathbf{X},\mathbf{Y}\in\mathcal{X},\mathcal{Y}} \mathrm{P}(\mathbf{Y},\mathbf{X})\ f(\mathbf{X},\mathbf{Y})\ \partial\mathbf{X}\ \partial\mathbf{Y}$. Just as in continuous control problems, which can only be solved exactly for carefully chosen formulations, the behavioral constraints, g_i and h_i, and the distribution of side information, $\mathrm{P}(\mathbf{X}^T||\mathbf{Y}^{T-1})$, must be specifically chosen to assure that the causally conditioned probability distribution, $P^*_{\lambda,\gamma}(\mathbf{Y}^T||\mathbf{X}^T)$, is a closed-form expression. We do not investigate the continuous-time setting in this work, though the ideas of this thesis are not restricted to discrete-time settings.

5.3.4 Deterministic Side Information

In the special case that $\mathrm{P}(\mathbf{X}^T||\mathbf{Y}^{T-1})$ is a deterministic distribution, the maximum causal entropy distribution simplifies significantly. Consider the bounds of the conditional and causal entropies in general and in the special deterministic case (Theorem 4.22). Thus, maximizing the causal entropy of $\mathbf{Y}$ given $\mathbf{X}$ is equivalent to maximizing the entropy of $\mathbf{Y}$ alone in the deterministic dynamics setting. Similarly, since the $\mathbf{X}$ variables can be fully recovered from the $\mathbf{Y}$ variables, the sets of behavioral constraint functions, g_i and h_j, can be expressed solely in terms of $\mathbf{Y}$ probability variables. Extending this to include static conditioning, $H(\mathbf{Y}^T||\mathbf{X}^T|\mathbf{W}) = H(\mathbf{Y}|\mathbf{W})$, and g_i and h_j can be expressed solely in terms of $\mathbf{W}$ and $\mathbf{Y}$ probability variables with no explicit dependence on $\mathbf{X}$ variables.

Maximizing the causal entropy of $\mathbf{Y}$ given $\mathbf{X}$ and statically conditioned on $\mathbf{W}$ in the deterministic setting reduces to a conditional entropy maximization:

$$\begin{aligned} &\underset{\mathrm{P}(\mathbf{Y}|\mathbf{W})}{\operatorname{argmax}} \; H(\mathbf{Y}^T|\mathbf{W}^T) \\ \text{such that: } &\forall_i \; g_i(\mathrm{P}(\mathbf{Y}|\mathbf{W})) = 0, \\ &\forall_j \; h_j(\mathrm{P}(\mathbf{Y}|\mathbf{W})) \leq 0, \\ &\forall_{\mathbf{W},\mathbf{Y}} \; \mathrm{P}(\mathbf{Y}|\mathbf{W}) \geq 0, \text{and} \\ &\forall_{\mathbf{W}} \; \sum_{\mathbf{Y}} \mathrm{P}(\mathbf{Y}|\mathbf{W}) - 1 = 0. \end{aligned} \tag{5.19}$$

The dual of this conditional entropy optimization (with probabilistic normalization constraints suppressed) is:

$$\inf_{\mathrm{P}(\mathbf{Y}|\mathbf{W})} \; -H(\mathbf{Y}|\mathbf{W}) + \sum_{i=1}^{m} \lambda_i g_i(\mathrm{P}(\mathbf{Y}|\mathbf{W})) + \sum_{j=1}^{n} \gamma_j h_j(\mathrm{P}(\mathbf{Y}|\mathbf{W})), \tag{5.20}$$

for $\gamma \geq 0$. As this special case no longer depends on the causal entropy, and only the conditional entropy, we shall show in Section 6.2.5 that for common choices of behavioral constraint functions this approach is consistent with previously developed probabilistic models—specifically a globally feature-conditioned variant of the conditional random field (Lafferty et al., 2001).

5.4 Discussion

The maximum causal entropy formulation presented in this chapter enables predictive models of behavior sequences that satisfy purposeful measures of empirical data (represented as inequality and equality constraints), while otherwise being as uncertain or agnostic as possible. Crucially, to match the information availability of many sequential data settings, the causal constraints we have introduced prevent the future side information variables from causally influencing past conditioned

variables. For predictive purposes, we have shown that the maximum causal entropy formulation provides important worst-case prediction guarantees—it minimizes the worst-case log-loss when information is revealed and predictions are made sequentially (Theorem 5.10). Additionally, the maximum causal entropy formulation is a convex optimization (Theorem 5.8), which provides efficiency benefits.

In this thesis, we are focused on modeling human behavior. However, it is important to note that the principle of maximum causal entropy is not specific to decision-making formalisms. Its predictive log-loss guarantees are applicable to any setting with sequentially revealed side information where there is no causal influence of side information over a predicted sequence of variables. We apply this general approach to behavior prediction tasks in Markov decision processes in Chapter 6, influence diagrams in Chapter 7, and game-theoretic equilibria in Chapter 8.

Chapter 6

Statistic-Matching Maximum Causal Entropy and Inverse Optimal Control

"Although nature commences with reason and ends in experience it is necessary for us to do the opposite, that is to commence with experience and from this to proceed to investigate the reason."

— Leonardo da Vinci (Scientist, 1452–1519).

We now introduce a specific form of maximum causal entropy distribution that matches the characteristics of observed data in probabilistic expectation (Ziebart et al., 2010b). This formulation is applicable for prediction in a wide range of problem settings characterized by information revelation, feedback, and stochastic interaction. However, we specifically consider the perspective of forecasting sequences of decisions in this chapter. Under this perspective, we interpret the Lagrange multipliers of the maximum causal entropy optimization as parameters for *utility potential functions* of a probabilistic decision-theoretic solution class, and address the inverse optimal control problem of recovering a utility function that provides performance guarantees and best explains demonstrated behavior.

The contributions of this chapter connect previously disparate concepts—inverse optimal control and exponential family probability distributions—for the first time. From the perspective of probabilistic graphical models, maximum causal entropy in the inverse optimal control setting generalizes conditional random fields (Lafferty et al., 2001) by allowing side information to be dynamically revealed from a known conditional probability distribution. From the control perspective, it augments the decision-theoretic performance guarantees of inverse optimal control with the information-theoretic prediction guarantees of the principle of maximum entropy.

6.1 Statistic Matching Constraints

6.1.1 General and Control Motivations and Problem Setting

When approximating a probability distribution, a common constraint for the approximation is that it match empirical statistics of the observed distribution (in expectation). In some domains, those empirical statistics are the only information available about a distribution due to a limited number of conducted experiments. In other domains, a selection of empirical statistics is assumed to capture the most salient properties of the distribution—for instance, selected based on background knowledge, such as physical significance. Matching those statistics (and not necessarily matching a large number of additional available statistics of the distribution) prevents overfitting to the empirical data and enables better predictions from smaller sample sizes. The maximum causal entropy approach extends the maximum entropy principle to be applied when some variables should not be causally influenced by others and matching empirical statistics of data is desired.

For control and decision-making domains, $\mathbf{Y}$ corresponds to the sequence of actions, $\mathbf{A}$, that an agent employs over time. The $\mathbf{X}$ variables correspond to the sequences of states, $\mathbf{S}$, of the agent over time. The dynamics governing the states are generally stochastic functions of previous states and actions, $\mathrm{P}(S_t|\mathbf{S}_{1:t-1}, \mathbf{A}_{1:t-1})$, and are often **Markovian**: $\mathrm{P}(S_t|S_{t-1}, A_{t-1})$. We assume that these dynamics are either explicitly provided or estimated from data using a separate procedure. Since future states are only revealed *after* actions are selected, they should have no causal influence over earlier actions. This matches the causal assumptions of the maximum causal entropy model of the conditional action distribution, $\{\mathrm{P}(A_t|\mathbf{A}_{1:t-1}, \mathbf{S}_{1:t})\}$. In this context, we refer to this conditional action distribution as a stochastic **policy**, denoted $\pi(A_t|\mathbf{A}_{1:t-1}, \mathbf{S}_{1:t})$, which, in the Markovian setting, reduces to $\pi(A_t|S_t)$. We refer to the statistics of the distribution to be matched in expectation as **feature functions**, $\mathcal{F}(\mathbf{S}, \mathbf{A}) \to \mathbb{R}^K$.

As we shall see, a number of interesting analogies to decision-theoretic models and inference procedures arise from this formulation. Namely, by interpreting the Lagrange multipliers of the maximum causal entropy model as utility function parameters, the probability distribution is inferred using a softened version of the Bellman equation (Bellman, 1957). This inherently captures purposefulness and adaptation in the model. However, it is in error to restrict applications of the maximum causal entropy approach only to data generated by mechanisms that are assumed to possess purposefulness and adaptation just because the model provides this interpretation. Indeed, though we focus on actions and states in our interpretation here, the predictive guarantees of the approach apply to any sequential data with causal influence restrictions and statistical features.

6.1.2 Optimization and Policy Form

We now formalize and solve the maximum causal entropy optimization for matching empirical expectations. Formally, the expected feature functions implied by the stochastic maximum causal entropy policy, $\mathrm{E}_{\mathrm{P}(\mathbf{S},\mathbf{A})}[\mathcal{F}(\mathbf{S}, \mathbf{A})]$, are constrained to match the empirical feature function expecta-

tions, $\mathrm{E}_{\tilde{\mathrm{P}}(\mathbf{S},\mathbf{A})}[\mathcal{F}(\mathbf{S},\mathbf{A})]$. This corresponds to affine equality constraint functions: $g_i(\mathrm{P}(\mathbf{A}||\mathbf{S})) = \mathrm{E}_{\mathrm{P}(\mathbf{S},\mathbf{A})}[\mathcal{F}_i(\mathbf{S},\mathbf{A})] - \mathrm{E}_{\tilde{\mathrm{P}}(\mathbf{S},\mathbf{A})}[\mathcal{F}_i(\mathbf{S},\mathbf{A})] = 0$.

Definition 6.1. *The general maximum causal entropy optimization (Equation 5.6) reduces to the following optimization problem:*

$$\begin{aligned} &\underset{\{\mathrm{P}(A_t|\boldsymbol{S}_{1:t},\boldsymbol{A}_{1:t-1})\}}{\operatorname{argmax}} \quad H(\boldsymbol{A}^T||\boldsymbol{S}^T) \\ \textit{such that: } &\mathrm{E}_{\mathrm{P}(\boldsymbol{S},\boldsymbol{A})}[\mathcal{F}(\boldsymbol{S},\boldsymbol{A})] = \mathrm{E}_{\tilde{\mathrm{P}}(\boldsymbol{S},\boldsymbol{A})}[\mathcal{F}(\boldsymbol{S},\boldsymbol{A})] \\ &\forall_{\boldsymbol{S}_{1:t},\boldsymbol{A}_{1:t}} \mathrm{P}(A_t|\boldsymbol{S}_{1:t},\boldsymbol{A}_{1:t-1}) \geq 0 \\ &\forall_{\boldsymbol{S}_{1:t},\boldsymbol{A}_{1:t-1}} \sum_{A_t} \mathrm{P}(A_t|\boldsymbol{S}_{1:t},\boldsymbol{A}_{1:t-1}) = 1. \end{aligned} \tag{6.1}$$

In Definition 6.1 we have expressed the objective and probability constraints in terms of the conditional probability factors of the causally conditioned probability. This factorization corresponds to the general factored maximum causal entropy dual in Equation 5.8. We note that as $\mathrm{P}(\mathbf{A},\mathbf{S}) = \mathrm{P}(\mathbf{A}^T||\mathbf{S}^T)\,\mathrm{P}(\mathbf{S}^T||\mathbf{A}^{T-1})$, the constraints of Equation 6.1 are affine in terms of $\mathrm{P}(\mathbf{A}^T||\mathbf{S}^T)$ variables, but not in terms of $\mathrm{P}(A_t|\mathbf{S}_{1:t},\mathbf{A}_{1:t-1})$ terms. When the feature mapping from state and action sequences to real-valued vectors is considered explicitly as a set of variables, the objective is $H(\mathbf{A}^T||\mathbf{S}^T|\{\mathcal{F}\})$. However, we will drop this additional static conditioning on $\{\mathcal{F}\}$ in our notation and formulation.

We now provide the factored form of the solution to this optimization problem.

Theorem 6.2. *The distribution satisfying the maximum causal entropy constrained optimization with feature function expectation constraints (Equation 6.1) has a form defined recursively as:*

$$\begin{aligned} \mathrm{P}_\theta(A_t|\boldsymbol{S}_{1:t},\boldsymbol{A}_{1:t-1}) &= \frac{Z_{A_t|\boldsymbol{S}_{1:t},\boldsymbol{A}_{1:t-1},\theta}}{Z_{\boldsymbol{S}_{1:t},\boldsymbol{A}_{1:t-1},\theta}} \\ \log Z_{\boldsymbol{S}_{1:t},\boldsymbol{A}_{1:t-1},\theta} &= \log \sum_{A_t} Z_{A_t|\boldsymbol{S}_{1:t},\boldsymbol{A}_{1:t-1},\theta} \\ &= \operatorname*{softmax}_{A_t} \Big(\sum_{S_{t+1}} \mathrm{P}(S_{t+1}|\boldsymbol{S}_{1:t},\boldsymbol{A}_{1:t}) \log Z_{\boldsymbol{S}_{1:t+1},\boldsymbol{A}_{1:t},\theta} \Big) \\ Z_{A_t|\boldsymbol{S}_{1:t},\boldsymbol{A}_{1:t-1},\theta} &= e^{\sum_{S_{t+1}} \mathrm{P}(S_{t+1}|\boldsymbol{S}_{1:t},\boldsymbol{A}_{1:t}) \log Z_{\boldsymbol{S}_{1:t+1},\boldsymbol{A}_{1:t},\theta}} \\ Z_{\boldsymbol{S}_{1:T},\boldsymbol{A}_{1:T-1},\theta} &= e^{\theta^\top \mathcal{F}(\boldsymbol{S},\boldsymbol{A})}, \end{aligned} \tag{6.2}$$

where $\operatorname{softmax}_x f(x) \triangleq \log \sum_x e^{f(x)}$.

Proof (sketch). We note that the (negated) primal objective function (Equation 6.1) is convex in the variables $\mathrm{P}(\mathbf{A}||\mathbf{S})$ and subject to linear constraints on feature function expectation matching, valid probability distributions, and non-causal influence of future side information (Theorem 5.8).

Differentiating the Lagrangian of the causal maximum entropy optimization (Equation 6.1), and equating to zero, we obtain the general form:

$$\mathrm{P}_\theta(A_t|\mathbf{S}_{1:t}, \mathbf{A}_{1:t-1}) \propto \exp\Big\{\theta^\top \mathrm{E}_{\mathrm{P}(\mathbf{S},\mathbf{A})}[\mathcal{F}(\mathbf{S},\mathbf{A})|\mathbf{S}_{1:t}, \mathbf{A}_{1:t}] - \sum_{\tau>t} \mathrm{E}_{\mathrm{P}(\mathbf{S},\mathbf{A})}[\log \mathrm{P}_\theta(A_\tau|\mathbf{S}_{1:\tau}, \mathbf{A}_{1:\tau-1})|\mathbf{S}_{1:t}, \mathbf{A}_{1:t}]\Big\}. \tag{6.3}$$

Substituting the more operational recurrence of Equation 6.2 into Equation 6.3 verifies the theorem. We will use this softmax interpretation (Equation 6.2) for the inference procedure extensively due to its close relation to decision-theoretic and optimal control inference procedures. □

The second exponentiated term in Equation 6.3 can be interpreted as a future expected causal entropy. It is difficult to directly interpret the meaning of the maximum causal entropy probability distribution from this general form. We provide a simplified version and relation to the Bellman equation in Section 6.2.1 after presenting a number of useful general properties of the distribution.

6.1.3 Properties

As a special case of the general maximum causal entropy approach, a number of important properties result from the statistic-matching constraint formulation.

Corollary 6.3 (of Theorem 5.10)**.** *The maximum causal entropy distribution minimizes the worst case prediction log-loss, i.e.,*

$$\inf_{\mathrm{P}(\boldsymbol{A}||\boldsymbol{S})} \sup_{\tilde{\mathrm{P}}(\boldsymbol{A}^T||\boldsymbol{S}^T)} -\sum_{\boldsymbol{A},\boldsymbol{S}} \tilde{\mathrm{P}}(\boldsymbol{A},\boldsymbol{S}) \log \mathrm{P}(\boldsymbol{A}^T||\boldsymbol{S}^T),$$

given $\tilde{\mathrm{P}}(\boldsymbol{A},\boldsymbol{S}) = \tilde{\mathrm{P}}(\boldsymbol{A}^T||\boldsymbol{S}^T)\,\mathrm{P}(\boldsymbol{S}^T||\boldsymbol{A}^{T-1})$ *and the constraint of matching feature expectations,* $\mathrm{E}_{\mathrm{P}(\boldsymbol{S},\boldsymbol{A})}[\mathcal{F}(\boldsymbol{S},\boldsymbol{A})] = \mathrm{E}_{\tilde{\mathrm{P}}(\boldsymbol{S},\boldsymbol{A})}[\mathcal{F}(\boldsymbol{S},\boldsymbol{A})]$*, when* $\boldsymbol{S}$ *is sequentially revealed from a known conditional probability distribution and actions are sequentially predicted using only previously revealed variables.*

Corollary 6.3 provides worst-case predictive log-loss guarantees for all possible distributions that match feature statistics.

As in other maximum entropy-based models, maximizing entropy is consistent with maximum likelihood estimation of model parameters–in this case, **maximum causal likelihood**.

Theorem 6.4. *Maximizing the causal entropy,* $H(\boldsymbol{A}^T||\boldsymbol{S}^T)$ *while constrained to match (in expectation) empirical feature functions,* $\tilde{\mathrm{E}}[\mathcal{F}(\boldsymbol{S},\boldsymbol{A})]$*, is equivalent to* ***maximum causal likelihood estimation*** *of* θ *given data set* $\{\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{S}}\}$ *under the conditional probability distribution of Equation 6.3:*

$$\begin{aligned}\hat{\theta} &= \operatorname*{argmax}_\theta \log \prod_i \mathrm{P}_\theta(\tilde{\boldsymbol{A}}^{(i)}||\tilde{\boldsymbol{S}}^{(i)}) \\ &= \operatorname*{argmax}_\theta \sum_{i,t} \log \mathrm{P}(\tilde{A}_t^{(i)}|\tilde{\boldsymbol{S}}_{1:t}^{(i)}, \tilde{\boldsymbol{A}}_{1:t-1}^{(i)}),\end{aligned}$$

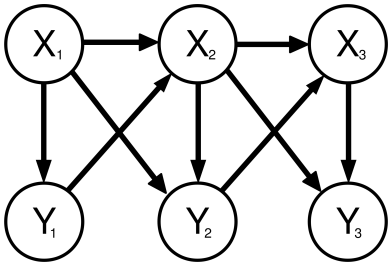

Figure 6.1: The Markovian factorization of **Y** variables causally conditioned on **X** variables.

where (i) *indexes the training examples.*

6.1.4 Markovian Simplifications

The maximum causal entropy policy of Equation 6.2 simplifies greatly when feature functions decompose over timesteps, *i.e.*, $\mathcal{F}(\mathbf{S}, \mathbf{A}) = \sum_t \mathbf{f}_{S_t, A_t}$, and side information dynamics are Markovian, *i.e.*, $\mathrm{P}(S_{t+1}|\mathbf{S}_{1:t}, \mathbf{A}_{1:t}) = \mathrm{P}(S_{t+1}|S_t, A_t)$.

Corollary 6.5 (of Theorem 6.2)**.** *The distribution satisfying the maximum causal entropy constrained optimization for the sequential problem setting of Equation 6.1 that is further assumed to have Markovian feature function expectation constraints,* $\mathcal{F}(\boldsymbol{S}, \boldsymbol{A}) = \sum_t \boldsymbol{f}_{S_t, A_t}$*, and Markovian side information dynamics,* $\mathrm{P}(S_{t+1}|S_t, A_t)$*, has a form defined recursively as:*

$$
\begin{aligned}
\mathrm{P}_\theta(A_t|S_t) &= \frac{Z_{A_t|S_t,\theta}}{Z_{S_t,\theta}} \qquad (6.4)\\
\log Z_{S_t,\theta} &= \log \sum_{A_t} Z_{A_t|S_t,\theta} \\
&= \operatorname*{softmax}_{A_t} \Big(\sum_{S_{t+1}} \mathrm{P}(S_{t+1}|S_t, A_t) \log Z_{S_{t+1},\theta} + \theta^\top \boldsymbol{f}_{S_t,A_t} \Big) \\
&= \operatorname*{softmax}_{A_t} \Big(\mathrm{E}_{\mathrm{P}(S_{t+1}|S_t,A_t)}[\log Z_{S_{t+1},\theta}] + \theta^\top \boldsymbol{f}_{S_t,A_t} \Big) \\
Z_{A_t|S_t,\theta} &= e^{\sum_{S_{t+1}} \mathrm{P}(S_{t+1}|S_t,A_t) \log Z_{S_{t+1},\theta} + \theta^\top \boldsymbol{f}_{S_t,A_t}} \\
&= e^{\mathrm{E}_{\mathrm{P}(S_{t+1}|S_t,A_t)}[\log Z_{S_{t+1},\theta}] + \theta^\top \boldsymbol{f}_{S_t,A_t}}
\end{aligned}
$$

where $\operatorname{softmax}_x f(x) \triangleq \log \sum_x e^{f(x)}$.

In general, any sequential set of variables can be "Markovianized" by augmenting the side information with variables that are sufficient statistics of the history of variables. Thus, this simplification and the perspective it provides is still quite general.

6.1.5 Goal-Directed Feature Constraints

A simple view of goal-directed behavior is that, when successful, it is characterized by a sequence of states and actions that terminate in an intended final state. Incorporating a terminal-state feature within the feature-matching maximum causal entropy framework (Equation 6.1) can constrain the policy distribution to terminate in the same goal states as demonstrated behavior.

There are many possible of goal-directed settings:

- **Unconstrained**: In the unconstrained setting, trajectories are either infinite or can terminate in any state. Actions are chosen only based on the uncertain future utility and are not constrained to reach particular states.

- **First exit**: There is a set of exit states, $\mathcal{S}_{\text{exit}}$, in which a sequence of states and actions can terminate. This can be accomplished in the maximum causal entropy framework by adding a feature to the maximum causal entropy optimization's constraints that is an indicator function of the terminal state: $\mathrm{E}_P[I(S_T \in \mathcal{S}_{\text{exit}})] = E_{\tilde{P}}[I(S_T \in \mathcal{S}_{\text{exit}})]$. The terminal state is then forced to match the empirical distribution: $\mathrm{P}(S_T \in \mathcal{S}_{\text{exit}}) = \tilde{\mathrm{P}}(S_T \in \mathcal{S}_{\text{exit}})$. If this probability is 1, only terminal states in $\mathcal{S}_{\text{exit}}$ are permitted.

- **Exit distribution**: The terminal state of state-action sequences is distributed according to $\tilde{\mathrm{P}}(\text{exit})$. This can be accomplished in the maximum causal entropy framework by augmenting the feature set with a K^{th} feature, $\mathcal{F}_K(\mathbf{S}_{1:T}, \mathbf{A}_{1:T-1}) = S_T$, and constraining that feature to match the empirical distribution in expectation: $\mathrm{P}(S_T) = \tilde{\mathrm{P}}(S_T)$.

- **Paired start-exit distribution**: The initial and terminal state pair is distributed according to the distribution $\tilde{\mathrm{P}}(\text{start}, \text{exit})$. This can be accomplished in the maximum causal entropy framework by augmenting the feature set with a K^{th} feature, $\mathcal{F}(\mathbf{S}_{1:T}, \mathbf{A}_{1:T-1}) = S_1 \times S_T$, and constraining that feature to match the empirical distribution in expectation: $\mathrm{P}(S_1, S_T) = \tilde{\mathrm{P}}(S_1, S_T)$.

The selection of a particular goal-directed setting is heavily application dependent. Additionally, for some behavior forecasting applications, behavior may be goal-constrained, but the particular terminal constraints may be unknown. We investigate a Bayesian approach to that problem in Chapter 11.

6.2 Inverse Optimal Control

Inverse optimal control, as described in Section 3.2.1, has traditionally been viewed as the task of recovering the reward function of a decision process (*e.g.*, a Markov decision process) that induces demonstrated behavior sequences. This is an important problem for programming robots by demonstration (Argall et al., 2009). We employ the statistic-matching maximum causal entropy approach to the problem of inverse optimal control. Applying the principle of maximum causal

entropy to this setting with statistic-matching constraints addresses the inverse optimal control problem with predictive guarantees. This approach connects inverse optimal control to exponential family probability distributions for the first time.

6.2.1 Policy Guarantees

We now specifically focus on the **inverse optimal control** problem of recovering the rewards of a parametric-reward Markov decision process (Definition 2.10), $\mathcal{M}_{\text{PRMDP}}/\theta$, that provides guarantees with respect to observed behavior. Previous approaches to the inverse optimal control problem suffer from a great deal of ill-posedness (Section 3.2.1, particularly Remark 3.2) that the maximum causal entropy approach resolves using its unique solution criteria and robustness to sub-optimal behavior.

Theorem 6.6 (from Abbeel & Ng (2004))**.** *Any distribution that matches feature function expectations,* $\mathrm{E}_{\mathrm{P}(\boldsymbol{S},\boldsymbol{A})}[\mathcal{F}(\boldsymbol{S},\boldsymbol{A})]$ *with demonstrated expectations,* $\mathrm{E}_{\tilde{\mathrm{P}}(\boldsymbol{S},\boldsymbol{A})}[\mathcal{F}(\boldsymbol{S},\boldsymbol{A})]$*, guarantees equivalent expected utility on the unknown parameters of a reward function linear in* $\mathcal{F}(\boldsymbol{S},\boldsymbol{A})$*.*

Corollary 6.7 (of Theorem 6.6 and Definition 6.1)**.** *The maximum causal entropy distribution provides the same expected utility as demonstrated variable sequences on any unknown parameters of a reward function linear in* $\mathcal{F}(\boldsymbol{S},\boldsymbol{A})$*.*

Theorem 6.6 and Corollary 6.7 show that the maximum causal entropy inverse optimal control approach provides the same utility matching guarantees as Abbeel & Ng (2004). However, unlike that past work, in addition to this utility-matching guarantee, the maximum causal entropy distribution also provides the strong worst-case predictive guarantee of Corollary 6.3. As a result, the model provides probabilistic support for all possible behaviors and non-zero probability for demonstrated behavior. This overcomes the disadvantages of previous approaches providing no support for demonstrated behavior illuminated by Remark 3.4.

6.2.2 Soft Bellman Equation Interpretation

Surprisingly, though we began with a formulation based on information theory, the resulting probability distribution has close connections to decision theory. In fact, the maximum causal entropy probability distribution is a generalization of optimal control laws. We now redefine the log partition functions of the simplified probability distribution (the lower Markovian-order version of Equation 6.2) in terms of a state-action value, $Q_\theta^{\text{soft}}(a_t, s_t)$, and a state-based value, $V_\theta^{\text{soft}}(s)$, to provide a useful analogy to the Bellman equation for determining optimal control and decision making. This connection enables straight-forward extension of the value iteration algorithm for solving the Bellman equation (Bellman, 1957) to the maximum causal entropy inverse optimal control setting.

Theorem 6.8. *The maximum causal entropy distribution with statistic matching (Theorem 6.2) can be re-expressed as:*

$$\begin{aligned} Q_\theta^{soft}(a_t, s_t) &\triangleq \log Z_{a_t|s_t} \\ &= \mathrm{E}_{\mathrm{P}(s_{t+1}|s_t,a_t)}[V_\theta^{soft}(s_{t+1})|s_t, a_t] + \theta^\top \boldsymbol{f}_{s_t,a_t} \qquad (6.5) \\ V_\theta^{soft}(s_t) &\triangleq \log Z_{s_t} \\ &= \operatorname*{softmax}_{a_t} Q_\theta^{soft}(a_t, s_t), \qquad (6.6) \end{aligned}$$

where $\operatorname{softmax}_x f(x) \triangleq \log \sum_x e^{f(x)}$ *provides a smooth interpolation (i.e., differentiable) of the maximum of different functions.*

We illustrate the behavior of the soft maximum function in Figure 6.2. Without loss of generality, when needed, the entire history of states and actions can be compressed into the current state, $S'_t = \{\mathbf{S}_{1:t}, \mathbf{A}_{1:t}\}$, making Equations 6.5 and 6.6 applicable to any "Markovianized" causal side information setting with feature constraints that linearly decompose over time[1].

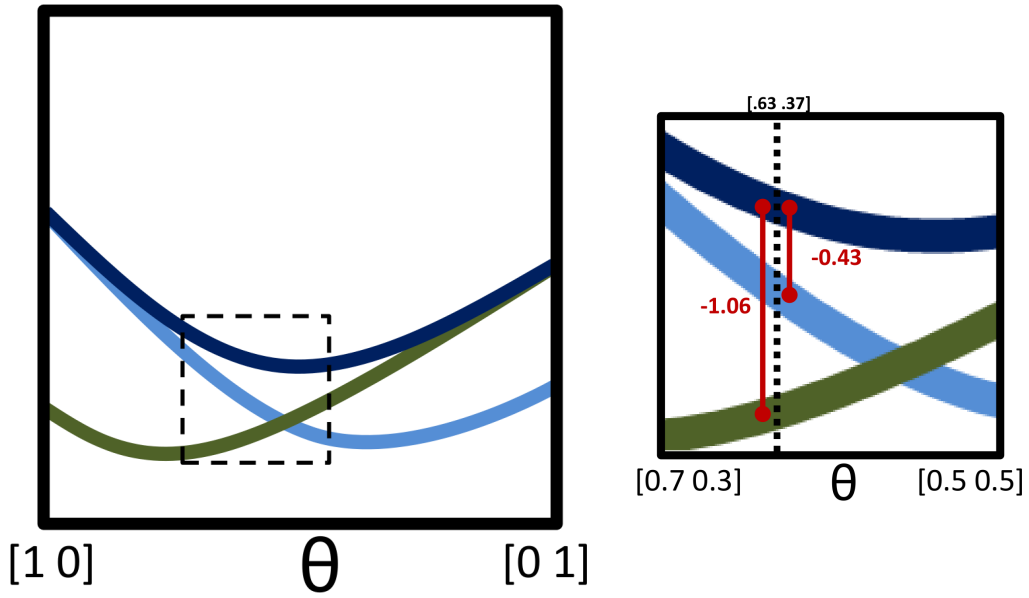

Figure 6.2: Left: Two action value functions, $Q^{\text{soft}}(a_1, s)$ and $Q_\theta^{\text{soft}}(a_2, s)$ and the soft maximum function, $V_\theta^{\text{soft}}(s) = \operatorname{softmax}_a Q_\theta^{\text{soft}}(a, s)$. Right: A zoomed in portion of the soft maximum function and the gaps between the action value function and the soft maximum function at a particular choice of parameters. Action probabilities are distributed according to the exponentiated gap distance.

The gap between an action's value and the state value, $Q_\theta^{\text{soft}}(s, a) - V_\theta^{\text{soft}}(s)$, determines that action's probability within the maximum causal entropy model: $\pi_\theta(a|s) = e^{Q_\theta^{\text{soft}}(s,a) - V_\theta^{\text{soft}}(s)}$. When

[1]This "Markovianization" is a common technique that is not unique to the approach of this thesis.

the gaps of multiple actions approach equivalence, the probability of those actions become uniform under the distribution. In the opposite limit, when the gap between one action and all others grows large, the stochastic maximum causal entropy policy becomes deterministic and matches the optimal policy (Remark 6.9).

Remark 6.9. *Let parameters, θ, be positively scaled by scalar α. As $\alpha \to \infty$, the soft maximum under scaled parameters, $\alpha\theta$, behaves like a maximum function and the maximum causal entropy recurrence relation (Equation 6.5 and Equation 6.6) becomes equivalent to the Bellman equation*[2].

We now analyze the maximum causal entropy distribution to establish a number of its desirable properties and relationships to optimal decision theory.

Theorem 6.10. *The probability of a stochastic policy, $\pi \triangleq \{\mathrm{P}(A_\tau | \boldsymbol{S}_{1:\tau}, \boldsymbol{A}_{1:\tau-1})\}$, under the maximum causal entropy distribution is related to the expected feature potentials and the softmax recurrence as follows:*

$$\log P_\theta^{soft}(\pi) = \mathrm{E}_{P(\boldsymbol{S}_{1:T}, \boldsymbol{A}_{1:T}|\pi)} \left[\sum_{t=1}^{T} \theta^\top \boldsymbol{f}_{S_t, A_t} \middle| \pi \right] - \sum_{S_1} \mathrm{P}(S_1) V_\theta^{soft}(S_1), \tag{6.7}$$

where the latter term is independent from the policy, π.

Theorem 6.10 establishes a monotonic relationship between policy probability and expected values under the soft maximum value iteration interpretation. This overcomes our earlier criticisms of other inverse optimal control approaches (Remark 3.4). Two desirable properties of the maximum causal entropy distribution follow immediately.

Corollary 6.11. *The most likely policy within the maximum causal entropy model maximizes the expected feature potentials:*

$$\pi^* = \operatorname*{argmax}_\pi \mathrm{E}_{\mathrm{P}(\boldsymbol{S}_{1:T}, \boldsymbol{A}_{1:T}|\pi)} \left[\sum_{t=1}^{T} \theta^\top \boldsymbol{f}_{S_t, A_t} \middle| \pi \right]. \tag{6.8}$$

These expected feature potentials can be interpreted as the expected value of the policy.

The correspondence between the most likely policy and maximizing expected feature potentials (Corollary 6.11) enables prescriptive solution techniques to be employed for finding the most likely policy within the maximum causal entropy probability distribution.

[2]When two action sequences provide equivalent reward, a uniform distribution over those entire sequences is prescribed under the maximum causal entropy model, while the Bellman equations arbitrarily choose one of the optimal actions to deterministically employ.

Corollary 6.12. *Two policies have equivalent probability within the maximum causal entropy model if and only if they have equal expected feature potential functions:*

$$P_\theta^{soft}(\pi_1) = P_\theta^{soft}(\pi_2) \Leftrightarrow \mathrm{E}_{\mathrm{P}(\boldsymbol{S}_{1:T},\boldsymbol{A}_{1:T}|\pi_1)}\left[\sum_{t=1}^{T}\theta^\top \boldsymbol{f}_{S_t,A_t}\right] = \mathrm{E}_{\mathrm{P}(\boldsymbol{S}_{1:T},\boldsymbol{A}_{1:T}|\pi_2)}\left[\sum_{t=1}^{T}\theta^\top \boldsymbol{f}_{S_t,A_t}\right]. \quad (6.9)$$

Corollary 6.12 shows that there exists no preference between policies with equivalent potential functions by the maximum causal entropy distribution. Thus, under the interpretation of the Lagrangian potentials as utility functions, an equal preference over equal expected utility policies exists.

6.2.3 Large Deviation Bounds

A natural question is: how many samples are needed to reasonably estimate the empirical feature counts of demonstrated behavior? Large deviation bound analysis enables us to bound the probability of sample feature counts from deviating greatly from the true distribution from which they are drawn.

Theorem 6.13. *The deviation between the empirical average of feature vectors and the expectation of feature vectors is bounded by:*

$$\mathrm{P}\left(\left\|\frac{1}{n}\sum_i \boldsymbol{F}_i - \mathrm{E}\left[\frac{1}{n}\sum_i \boldsymbol{F}_i\right]\right\|_\infty \geq \epsilon\right) \leq \sum_{k=1}^{K} 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(f_{i,k}^{max} - f_{i,k}^{min})^2}\right), \quad (6.10)$$

via Hoeffding's inequality and the union bound, where $\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots$ are random variables corresponding to expected feature vectors obtained by policies (random variables), and assuming that those feature samples are bounded by $P(F_{i,k} - E[F_{i,k}] \in [f_{i,k}^{min}, f_{i,k}^{max}]) = 1$. In the special case that all elements of the difference of sampled feature vectors from their expectation are bounded by the same values, this reduces to:

$$\mathrm{P}\left(\left\|\frac{1}{n}\sum_i \boldsymbol{F}_i - \mathrm{E}\left[\frac{1}{n}\sum_i \boldsymbol{F}_i\right]\right\|_\infty \geq \epsilon\right) \leq 2K\exp\left(-\frac{2n\epsilon^2}{(f^{max} - f^{min})^2}\right). \quad (6.11)$$

When the reward function parameters are known, a similar question can be asked about the average reward of randomly sampled policies. This question is answered by Theorem 6.14.

Theorem 6.14. *The deviation between the empirical average reward and the expected reward is bounded by:*

$$\mathrm{P}\left(\left|\frac{1}{n}\sum_i \theta^\top \boldsymbol{F}_i - \mathrm{E}\left[\frac{1}{n}\sum_i \theta^\top \boldsymbol{F}_i\right]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(reward_i^{max} - reward_i^{min})^2}\right), \quad (6.12)$$

where $\theta^\top \boldsymbol{F}_1, \theta^\top \boldsymbol{F}_2, ...,$ are the expected rewards obtained from policies with expected features $\boldsymbol{F}_1, ...$ under policies (random variables), and assuming that the rewards are bounded by $P(\theta^\top \boldsymbol{F}_i - E[\theta^\top \boldsymbol{F}_i] \in [\mathit{reward}_i^{\mathit{min}}, \mathit{reward}_i^{\mathit{max}}]) = 1$. In the special case that the bounds on the rewards are the same, this reduces to:

$$\mathrm{P}\left(\left|\frac{1}{n}\sum_i \theta^\top \boldsymbol{F}_i - \mathrm{E}\left[\frac{1}{n}\sum_i \theta^\top \boldsymbol{F}_i\right]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2n\epsilon^2}{(\mathit{reward}^{\mathit{max}} - \mathit{reward}^{\mathit{min}})^2}\right). \tag{6.13}$$

6.2.4 Continuous Problems with Linear Dynamics and Quadratic Utilities

Often in problems of control (*e.g.*, of a robot or aircraft), the number of states and actions are infinite and correspond to continuous-valued quantities. One application of the principle of maximum entropy to this domain is to constrain quadratic and linear state expectations[3]:

$$\mathrm{E}_{\mathrm{P}}\left[\sum_t \mathbf{s}_t\, \mathbf{s}_t^\top\right] = \mathrm{E}_{\tilde{\mathrm{P}}}\left[\sum_t \mathbf{s}_t\, \mathbf{s}_t^\top\right]; \text{ and}$$
$$\mathrm{E}_{\mathrm{P}}\left[\sum_t \mathbf{s}_t\right] = \mathrm{E}_{\tilde{\mathrm{P}}}\left[\sum_t \mathbf{s}_t\right].$$

These constraints yield Lagrangian terms: $\mathbf{Q} \cdot \sum_t \mathbf{s}_t\, \mathbf{s}_t^\top$ and $\mathbf{R} \cdot \sum_t \mathbf{s}$, where $\mathbf{Q}$ is a matrix and $\mathbf{R}$ is a vector[4]. These Lagrangian potentials can be equivalently expressed as: $\sum_t \mathbf{s}_t^\top \mathbf{Q} \mathbf{s}_t$ and $\sum_t \mathbf{s}_t^\top \mathbf{R}$.

The softened Bellman equations for this settings are then as follows:

$$\begin{aligned} Q_\theta^{\text{soft}}(\mathbf{a}_t, \mathbf{s}_t) &= \mathrm{E}_{\mathrm{P}(\mathbf{s}_{t+1}|\mathbf{s}_t,\mathbf{a}_t)}[V_\theta^{\text{soft}}(\mathbf{s}_{t+1})|\mathbf{s}_t, \mathbf{a}_t] \\ V_\theta^{\text{soft}}(\mathbf{s}_t) &= \operatorname*{softmax}_{\mathbf{a}_t} Q_\theta^{\text{soft}}(\mathbf{a}_t, \mathbf{s}_t) + \mathbf{s}_t^\top \mathbf{Q} \mathbf{s}_t + \mathbf{s}_t^\top \mathbf{R}, \end{aligned} \tag{6.14}$$

where $\mathbf{Q}$ is a positive semi-definite matrix and $\mathrm{P}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$ is distributed according to a multivariate conditional Gaussian. Based on the equivalence between the form of a Gaussian distribution and an exponentiated quadratic function, iterative computation of the value function is possible in closed form (Theorem 6.15).

Theorem 6.15. *For the special case where dynamics are linear functions with Gaussian noise, the quadratic MaxCausalEnt model permits a closed-form solution and, given dynamics $\boldsymbol{s}_{t+1} \sim$*

[3] Quadratic and linear constraints on the action variables, $E_{\mathrm{P}}[\sum_t \mathbf{s}_t \mathbf{s}_t^\top] = E_{\tilde{\mathrm{P}}}[\sum_t \mathbf{s}_t \mathbf{s}_t^\top]$ and $E_{\mathrm{P}}[\sum_t \mathbf{s}_t^\top] = E_{\tilde{\mathrm{P}}}[\sum_t \mathbf{s}_t^\top]$. However, as described in Section 2.2.2, augmenting the state with past action values enables the equivalent constraints solely in terms of state constraints.

[4] We employ the Frobenius inner product (also referred to as the matrix dot product) for the first Lagrangian potential term: $\mathbf{A} \cdot \mathbf{B} = \sum_{i,j} \mathbf{A}_{i,j} \mathbf{B}_{i,j}$ and the standard vector dot product for the second Lagrangian potential term.

$N(\boldsymbol{A}\boldsymbol{s}_t + \boldsymbol{B}\boldsymbol{a}_t, \Sigma)$*, Equation 2.10 reduces to:*

$$Q_\theta^{soft}(\boldsymbol{a}_t, \boldsymbol{s}_t) = \begin{bmatrix} \boldsymbol{a}_t \\ \boldsymbol{s}_t \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{B}^\top \boldsymbol{D}\boldsymbol{B} & \boldsymbol{A}^\top \boldsymbol{D}\boldsymbol{B} \\ \boldsymbol{B}^\top \boldsymbol{D}\boldsymbol{A} & \boldsymbol{A}^\top \boldsymbol{D}\boldsymbol{A} \end{bmatrix} \begin{bmatrix} \boldsymbol{a}_t \\ \boldsymbol{s}_t \end{bmatrix} + \begin{bmatrix} \boldsymbol{a}_t \\ \boldsymbol{s}_t \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{B}^\top \boldsymbol{G} \\ \boldsymbol{A}^\top \boldsymbol{G} \end{bmatrix}$$

$$V_\theta^{soft}(\boldsymbol{s}_t) = \boldsymbol{s}_t^\top (\boldsymbol{C}_{s,s} + \boldsymbol{Q} - \boldsymbol{C}_{a,s}^\top \boldsymbol{C}_{a,a}^{-1} \boldsymbol{C}_{a,s}) \boldsymbol{s}_t + \boldsymbol{s}_t^\top (\boldsymbol{F}_s + \boldsymbol{R}) + const,$$

where $\boldsymbol{C}$ *and* $\boldsymbol{D}$ *are recursively computed as:* $\boldsymbol{C}_{a,a} = \boldsymbol{B}^\top \boldsymbol{D}\boldsymbol{B}; \boldsymbol{C}_{s,a} = \boldsymbol{C}_{a,s}^\top = \boldsymbol{B}^\top \boldsymbol{D}\boldsymbol{A}; \boldsymbol{C}_{s,s} = \boldsymbol{A}^\top \boldsymbol{D}\boldsymbol{A}; \boldsymbol{D} = \boldsymbol{C}_{s,s} + \boldsymbol{Q} - \boldsymbol{C}^\top \boldsymbol{C}_{a,a}^{-1} \boldsymbol{C}_{a,s}$*; and* $\boldsymbol{G} = \boldsymbol{F}_s + \boldsymbol{R}$*.*

The maximum causal entropy policy for the linear-quadratic setting is then distributed according to: $P(\mathbf{a}_t | \mathbf{s}_t) \propto e^{Q_\theta^{\text{soft}}(\mathbf{a}_t, \mathbf{s}_t)}$, which is also Gaussian.

6.2.5 Deterministic Dynamics Reduction

The inverse optimal control problem generalizes existing techniques for maximum entropy. The conditional random field (Lafferty et al., 2001) is a common example. When the conditional distribution of side information (*i.e.*, the state transition dynamics) is deterministic, the causal entropy (with static conditioning on features, $\{f_{s,a}\}$) reduces to the conditional entropy (Theorem 4.22). The optimization of Equation 6.1 reduces in this setting to the following optimization:

$$\max_{\mathbf{A}} H(\mathbf{A} | \{\mathbf{f}_{s,a}\}) \tag{6.15}$$

$$\text{such that: } \mathrm{E}_{\mathrm{P}(\mathbf{A})} \left[\sum_t \mathbf{f}_{s_t, a_t} \right] = \mathrm{E}_{\tilde{\mathrm{P}}(\mathbf{A})} \left[\sum_t \mathbf{f}_{s_t, a_t} \right]$$

$$\sum_{\mathbf{A}} \mathrm{P}(\mathbf{A}) = 1.$$

The form of the distribution over action sequences satisfying this optimization is:

$$\mathrm{P}(\mathbf{A}) \propto e^{\sum_t \theta^\top \mathbf{f}_{s_t, a_t}}. \tag{6.16}$$

When we consider the characteristics, $\{\mathbf{f}_{s,a}\}$, as variables on which the distribution is conditioned, this probability distribution corresponds to a conditional random field (Lafferty et al., 2001) where each A_t variable is conditioned on all characteristics, $\{\mathbf{f}_{s,a}\}$ (though it only depends on the set of features associated with the value it takes on). This application of conditional random fields differs from typical applications, which are typically concerned with recognition of a sequence of hidden variables from a sequence of noisy observations.

6.3 Relation to Alternate Information-Theoretic Approaches

A natural question in the statistic-matching/inverse optimal control setting is: what benefits does the principle of maximum *causal* entropy provide compared to existing maximum joint, marginal,

and conditional entropy approaches? We investigate this question in this section by exploring maximum entropy approaches based on previous information-theoretic concepts.

6.3.1 Maximum Joint Entropy

A natural approach to attempt is to maximize the joint entropy of $\mathbf{S}$ and $\mathbf{A}$. We note the decomposition of the joint distribution into two causal distributions relates the joint entropy to the causal entropy (Remark 4.21). This decomposition shows the difference between these two measures—the extra causally conditioned side information entropy, $H(\mathbf{S}^T||\mathbf{A}^{T-1})$:

$$H(\mathbf{A}, \mathbf{S}) = H(\mathbf{A}^T||\mathbf{S}^T) + H(\mathbf{S}^T||\mathbf{A}^{T-1}). \tag{6.17}$$

A simple application of the maximum joint entropy approach is to employ the Markov random field joint distribution,

$$\mathrm{P}(\mathbf{A}, \mathbf{S}) \propto e^{\theta^\top \mathcal{F}(\mathbf{S},\mathbf{A})}, \tag{6.18}$$

and obtain a conditional probability distribution by marginalizing the latent future state variables of the joint distribution:

$$\mathrm{P}(A_t|\mathbf{S}_{1:t}, \mathbf{A}_{1:t-1}) \propto \sum_{S_{t+1:T}, A_{t+1:T}} e^{\theta^\top \mathcal{F}(\mathbf{S},\mathbf{A})}.$$

Unfortunately this probability distribution is inefficient to compute and it ignores the crucial dependence between sequences with shared prefixes, which prevents two similar sequences from having probability distributions that are independent from one another.

6.3.2 Causally-Constrained Maximum Joint Entropy

A more sophisticated approach employs the causal constraints of the maximum causal entropy formulation on the joint entropy measure. The joint entropy is maximized while constraining the joint probability distribution to factor into the unknown causally conditioned probability of $\mathbf{A}$ given $\mathbf{S}$, $\mathrm{P}(\mathbf{A}^T||\mathbf{S}^T)$, and the known causal conditional probability of $\mathbf{S}$ given $\mathbf{A}$, $\mathrm{P}(\mathbf{S}^T||\mathbf{A}^{T-1})$, as follows:

$$\begin{aligned} &\underset{\{\mathrm{P}(A_t|\mathbf{S}_{1:t},\mathbf{A}_{1:t-1})\}}{\operatorname{argmax}} \quad H(\mathbf{A}, \mathbf{S}) \\ \text{such that: } &\mathrm{E}_{\mathrm{P}(\mathbf{S},\mathbf{A})}[\mathcal{F}(\mathbf{S}, \mathbf{A})] = \mathrm{E}_{\tilde{\mathrm{P}}(\mathbf{S},\mathbf{A})}[\mathcal{F}(\mathbf{S}, \mathbf{A})] \\ &\forall_{\mathbf{S}_{1:t},\mathbf{A}_{1:t}} \mathrm{P}(A_t|\mathbf{S}_{1:t}, \mathbf{A}_{1:t-1}) \geq 0; \text{and} \\ &\forall_{\mathbf{S}_{1:t},\mathbf{A}_{1:t-1}} \sum_{A_t} \mathrm{P}(A_t|\mathbf{S}_{1:t}, \mathbf{A}_{1:t-1}) = 1. \end{aligned} \tag{6.19}$$

The form of the conditional probability distribution satisfying this optimization is then:

$$\mathrm{P}_\theta(A_t|\mathbf{S}_{1:t}, \mathbf{A}_{1:t-1}) \propto \exp\Big\{\theta^\top \mathrm{E}_{\mathrm{P}(\mathbf{S},\mathbf{A})}[\mathcal{F}(\mathbf{S},\mathbf{A})|\mathbf{S}_{1:t}, \mathbf{A}_{1:t}] - \sum_{\tau>t} \mathrm{E}_{\mathrm{P}(\mathbf{S},\mathbf{A})}[\log \mathrm{P}_\theta(A_\tau|\mathbf{S}_{1:\tau}, \mathbf{A}_{1:\tau-1}) + \log \mathrm{P}(S_\tau|\mathbf{S}_{1:\tau-1}, \mathbf{A}_{1:\tau-1})|\mathbf{S}_{1:t}, \mathbf{A}_{1:t}]\Big\}. \tag{6.20}$$

Relating this distribution to the Bellman equation (Bellman, 1957), we have:

$$\begin{aligned} Q^{\mathrm{CCJE}}(a_t, s_t) &\triangleq \mathrm{E}_{\mathrm{P}(s_{t+1}|s_t,a_t)}[V^{\mathrm{CCJE}}(s_{t+1})|s_t, a_t] + \theta^\top \mathbf{f}_{s_t,a_t} + H(s_{t+1}|s_t, a_t) \\ V^{\mathrm{CCJE}}(s_t) &\triangleq \operatorname*{softmax}_{a_t} Q^{\mathrm{CCJE}}(a_t, s_t). \end{aligned} \tag{6.21}$$

The key difference from the maximum causal entropy soft-maximum interpretation is the inclusion of the entropy associated with the state transition dynamics, $H(s_{t+1}|s_t, a_t)$. Including this entropy can only decrease the uncertainty over action sequences. We argue that this is inappropriate when the dynamics of side information, $\mathrm{P}(\mathbf{S}^T||\mathbf{A}^{T-1})$, are known, or simply not variables of interest in the prediction task.

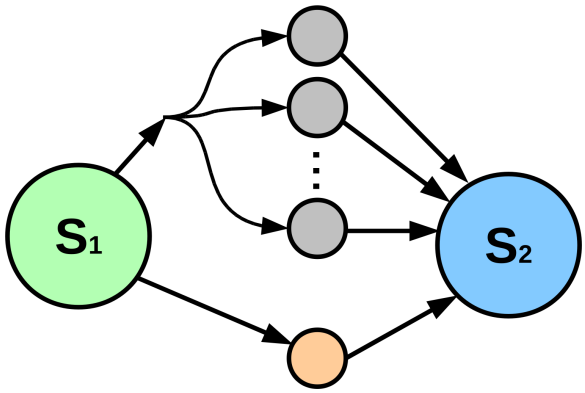

Figure 6.3: A Markov decision process that illustrates the implications of considering the uncertainty of stochastic dynamics under the causally conditioned maximum joint entropy model. Two actions are available in state S_1: one that has a high degree of next state uncertainty (top) and one that has a deterministic next state (bottom).

Figure 6.3 illustrates the sensitivity of the action distribution to the entropy of state transition dynamics. As the number of stochastic next states in the top action increases, the action distribution under the causally constrained maximum joint entropy is biased heavily towards that top action. However, this causally conditioned joint entropy could be appropriate for settings where both conditioned and side information distributions are being estimated. We do not investigate that setting in detail in this thesis.

6.3.3 Marginalized Maximum Conditional Entropy

We previously showed how the conditional random field can be derived by maximizing the conditional entropy of a probability distribution subject to feature-matching constraints in Section 5.1.3. The resulting probability distribution is of the form:

$$\mathrm{P}(\mathbf{A}|\mathbf{S}) \propto e^{\sum_t \theta^\top \mathbf{f}_{a_t,s_t}}. \tag{6.22}$$

Since the future state variables, $\mathbf{S}_{t+1:T}$, are unavailable when A_t is predicted, a practical approach is to take the joint distribution using Equation 6.22 and the transition dynamics and marginalize over the latent future state:

$$\begin{aligned}\mathrm{P}(a_t|s_t) &\propto \sum_{\mathbf{s}_{t+1:T}} \prod_{\tau>t} \mathrm{P}(s_{\tau+1}|s_\tau, a_\tau) e^{\sum_\tau \theta^\top \mathbf{f}_{a_\tau,s_\tau}} \\ &= \sum_{\mathbf{s}_{t+1:T}} e^{\sum_\tau \theta^\top \mathbf{f}_{a_\tau,s_\tau} + \log \mathrm{P}(s_{\tau+1}|s_\tau,a_\tau)}\end{aligned} \tag{6.23}$$

Theorem 6.16. *The marginalized maximum conditional entropy distribution (Equation 6.23) can then be interpreted under the dynamic programming perspective using the re-expression:*

$$\begin{aligned}Q_\theta^{CE}(a_t, s_t) &= \operatorname*{softmax}_{s_{t+1}} \left(V_\theta^{CE}(s_{t+1}) + \log \mathrm{P}(s_{t+1}|s_t, a_t)\right) + \theta^\top \boldsymbol{f}_{a_t,s_t} \\ V_\theta^{CE}(s_t) &= \operatorname*{softmax}_{a_t} Q_\theta^{CE}(a_t, s_t)\end{aligned} \tag{6.24}$$

If we view the Lagrangian potential function, $\theta^\top \mathbf{f}_{a_t,s_t}$, as a reward, then under this model an additional cost of $-\log \mathrm{P}(S_{t+1}|\mathbf{S}_{1:t}, \mathbf{A}_{1:t})$ is "paid" to obtain the desired outcome from the transition dynamics. As a result, the recovered policy is not equivalent to the corresponding Markov decision process' optimal policy.

Remark 6.17 (based upon Toussaint (2009), Equation 12)**.** *By Jensen's inequality, the log likelihood of the conditional random field distribution (Equation 3.2) corresponds to an upper bound on the expected reward (i.e., negative expected cost) in the corresponding MDP rather than the exact expected reward:*

$$\begin{aligned}\log \mathrm{E}_{\boldsymbol{S},\boldsymbol{A}}\left[\prod_{t=0}^{T} e^{-\mathit{cost}(A_t,S_t)} \middle| \pi(A|S)\right] &\geq \mathrm{E}_{\boldsymbol{S},\boldsymbol{A}}\left[\log \prod_{t=0}^{T} e^{-\mathit{cost}(A_t,S_t)} \middle| \pi(A|S)\right] \\ &= \mathrm{E}_{\boldsymbol{S},\boldsymbol{A}}\left[-\sum_{t=0}^{T} \mathit{cost}(A_t, S_t) \middle| \pi(A|S)\right].\end{aligned} \tag{6.25}$$

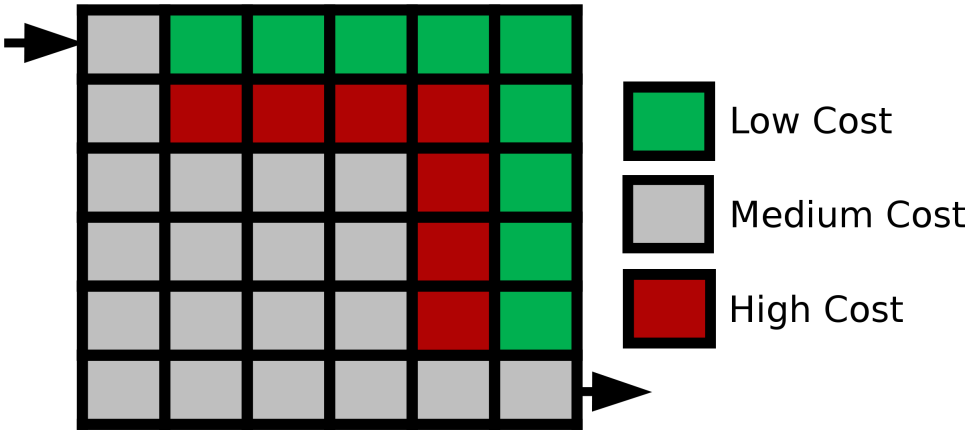

Figure 6.4: A Markov decision process that illustrates the preference of the marginalized maximum conditional entropy for "risky" actions that have small probabilities of realizing high rewards. Consider actions in the four cardinal directions from each state (*i.e.*, grid cell), and transition dynamics that with some probability of failure, a transition is made to an orthogonal cardinal direction instead of the intended direction. Green cells have very low cost, gray cells have medium, and red cells have extremely high cost.

The relationship between these two objectives is shown in Remark 6.17. Thus, the model has a strong preference for sequences of actions where the expected utility obtainable by any actual policy is quite low (due to the stochastic transition dynamics leading to low-reward outcomes), but a few "lucky" sequences have high utility.

This preference for risky state dynamics is illustrated in Figure 6.4, where a high utility trajectory to a goal state is possible, but unlikely due to stochastic state transition dynamics. Those dynamics make an alternate policy—going "around" rather than taking the risky path—have higher expected utility, but that policy is less preferred under the marginal maximum conditional entropy model, which assumes that desirable state transitions can be realized by paying a penalty.

6.4 Discussion

In this chapter, we have derived the probability distribution of the maximum causal entropy distribution for the statistic-matching constraint setting. We have shown that the problem of inverse optimal control can be posed in this way, and that the Lagrangian potentials, $\theta^\top \mathbf{f}$, that result can be interpreted as analogs to rewards or costs in existing decision frameworks. This connection is significant: prediction of purposeful behavior follows the same mechanisms as optimal generation approaches for that behavior (with added relaxations to address uncertainty). It is also surprising: after beginning with a purely information-theoretic formulation of an estimation problem, we have arrived at a stochastic generalization of optimal control criteria.

Other work has similarly aimed at combining information theory with decision theory and

warrants discussion. One prevalent line of research is determining the value of information, which is the amount an agent in a decision framework would be willing to pay to observe the value of a variable before making a decision (Howard, 1966; Poh & Horvitz, 1996; Krause & Guestrin, 2005). This perspective is useful for prescriptive purposes, but not for the predictive task we address with the maximum causal entropy approach. Tishby & Polani (2010) define an *information-to-go* quantity as an information-theoretic analog of the decision-theoretic *cost-to-go*. It is a measure of the future uncertainty of states and actions in decision settings, relative to a prior belief about the marginal distributions of future states and actions. This is similar to the causally-constrained maximum joint entropy approach and suffers the same bias for highly uncertain state transition dynamics. However, it may be useful if predicting both the future actions and states is desired.

Chapter 7

Maximum Causal Entropy Influence Diagrams for Imperfect Information

"Real knowledge is to know the extent of one's ignorance."
— Confucius (Philosopher, 551–479BC).

In the previous chapter, we investigated reasoning in the presence of future uncertainty. Many applications of machine learning and decision making also involve reasoning in the presence of historic and current uncertainty. For example, in partially observable Markov decision processes (POMDPs), only noisy observations of the world's state are revealed to the agent, and the agent is forced to act based on an uncertain belief about the world. This chapter extends the maximum causal entropy approach to settings with side information that includes latent variables and problems with imperfect recall where past observations are "forgotten." We employ a maximum causal entropy variant of the influence diagram (Miller et al., 1976; Howard & Matheson, 1984), a graphical framework that subsumes Bayesian Networks and augments their capabilities to reason about latent variables with *decisions* and *utilities* so that inference includes optimal decision making.

Maximum causal entropy expands the applicability of influence diagrams from the *prescription* of optimal decisions to the *prediction* of decision-making and the recovery of explanatory utility weights from observed decision sequences. Additionally, influence diagrams provide a convenient graphical representation of variable relationships in behavior prediction tasks. We investigate the structural properties of influence diagrams required for the maximum causal entropy approach to be applicable, finding that perfect decision recall is required.

7.1 Maximum Causal Entropy Influence Diagrams

We begin by describing different types of imperfect information and examples for each to motivate the contributions of this chapter before introducing the maximum causal entropy influence diagram.

7.1.1 Imperfect Information

Supporting the additional uncertainty of latent variables from the current point in time and from the past extends the maximum causal entropy model to behavior prediction in partial information, multi-player, and bounded rationality settings. For example, in many multi-player games (*e.g.*, card games such as poker), each player possesses a partial and distinctive knowledge of the state of the game and acts by inferring the other players' knowledge. Explicitly modeling this dynamic information availability is important for the prediction of actions in such settings.

We consider four different types of imperfect information in sequential decision settings:

1. **Latent future side information:** Future side information variables, $\mathbf{X}_{t+1:T}$, are unavailable when each conditioned variable, Y_t, is chosen. Example: the maximum causal entropy inverse optimal control setting of Chapter 6 (*e.g.*, control of a system in settings with stochastic dynamics).

2. **Partially observed side information:** Side information variables, $\mathbf{X}_{1:t}$, are only partially observed at each time step. Example: behavior in a partially observable Markov decision process (*e.g.*, the exploratory diagnosis of a partially-observed system).

3. **Latent previously observed side information:** Conditioning variables from the parent set of an early conditioned variable, par(Y_{t_1}), that are relevant to a future conditioned variable, Y_{t_2} ($t_1 < t_2$), but "forgotten" and not part of the parent set of Y_{t_2}. Example: games where players have secret information (*e.g.*, simple variants of poker).

4. **Latent past decisions:** Earlier conditioned variables (Y_{t_1}), that are relevant to future conditioned variables, Y_{t_2} ($t_1 < t_2$), but "forgotten" and not part of the parent set of Y_{t_2}. Example: multi-player games with hidden (or simultaneous) actions in otherwise full information settings.

In this chapter, we discuss the applicability of the maximum causal entropy approach to the latter three information settings. We generalize the dependency structure of maximum causal entropy inverse optimal control from Chapter 6 to address the second and third imperfect information settings.

7.1.2 Representation: Variables, Dependencies, and Features

We employ the influence diagram graphical model structure as a convenient representation for the estimated variables ($\mathbf{Y}$), the conditioning variables ($\mathbf{X}$), and the feature functions, $\mathcal{F}(\mathbf{X}, \mathbf{Y})$. We again employ the decision-based interpretation of these variables, replacing $\mathbf{Y}$ with actions, $\mathbf{A}$, and $\mathbf{X}$ with state variables, $\mathbf{S}$.

The maximum causal entropy influence diagram differs from the traditional influence diagram (Section 2.2.3) primarily in the meaning of utilities.

Definition 7.1. *A maximum causal entropy influence diagram (MaxCausalEnt ID) is structurally characterized by three different types of nodes and directed edges connecting those nodes. The node types are:*

- *Square decision nodes (**A**) that correspond to conditioned variables;*
- *Circular uncertainty nodes (**S**) that correspond to observed and unobserved side information variables; and*
- *Diamond utility nodes (**U**) that correspond to statistic-matching potential functions.*

Additionally, directed edges connect these nodes with the role of each edge depending on the type of node to which it is a parent[1]*. The roles are as follows:*

- *The parents of a decision node, par(A_t), are conditioned on or "known" when the A_t variable is assigned an instantiation value.*
- *An uncertainty node's parents, $par(S_t)$, specify the variables upon which its conditional probability distribution, $\mathrm{P}(S_t|par(S_t))$, is defined.*
- *The parents of a utility node, $par(U_t)$, in a MaxCausalEnt ID indicate the form of* feature functions *of the utility node $\mathcal{F}_{U_t} : par(U_t) \to \mathbb{R}^k$.*

The variables and relationships of the maximum causal entropy influence diagrams are summarized in Table 7.1.

Table 7.1: Influence diagram graphical representation structural elements, symbols, and relationships.

Type	Symbol	Parent relationship
Decision nodes	A	Specifies the variables that are observed when action A is selected
Uncertainty nodes	S	Specifies the conditional probability distribution, $\mathrm{P}(S\|\mathrm{par}(S))$, of the state variable S
Utility nodes	U	Specifies feature utility functions, $\theta^\top F(\mathrm{par}(U)) \to \mathbb{R}$, with (unknown) parameter weights θ

The only requirement on the structure of a maximum causal entropy influence diagram is acyclicity. However, the complexity of reasoning in the MaxEnt ID depends greatly on the dependencies between decision distributions implied by the structure (*i.e.*, the other decision distributions that are required to infer a particular decision node's distribution). Instead, those problems can often be more efficiently solved by exploiting the factored distribution of the side information.

[1] We assume utility nodes are not parents of any other node, but the same desired relationship—knowing what utility is received—can be accomplished by employing an intermediary uncertainty node that takes on the utility.

7.2 Maximum Causal Parent Entropy

Using the structure of the influence diagram, we now formulate the corresponding maximum causal entropy optimization, the structural constraints required by the formulation, and the probability distribution form.

7.2.1 Formulation and Optimization

We formulate the imperfect information maximum causal entropy optimization problem by allowing a distribution over side information variables, $\mathbf{X}$, and conditioned variables, $\mathbf{Y}$, that forms a joint distribution as:

$$\mathrm{P}(\mathbf{Y}, \mathbf{X}) = \prod_t \mathrm{P}(Y_t | \mathrm{par}(Y_t)) \prod_t \mathrm{P}(X_t | \mathrm{par}(X_t)), \tag{7.1}$$

where $\mathrm{par}(Y_t) \subseteq (Y_{1:t-1} \cup \mathbf{X})$ represents a set of parent variables of Y_t, and $\mathrm{par}(X_t)$ are constrained so that the directed graph formed by edges $\{e_{\mathrm{par}(Y_t) \to Y_t}\} \cup \{e_{\mathrm{par}(X_t) \to X_t}\}$ is acyclic.

Extending the setting of Chapter 5, we assume that the conditional distribution of side information, $\mathrm{P}(\mathbf{X} || \mathrm{par}(\mathbf{X}))$, is known, and we are left to estimate the conditional distribution, $\mathrm{P}(\mathbf{Y} || \mathrm{par}(\mathbf{Y}))$.

The causally conditioned probability distribution and entropy measure extend to this influence diagram graphical structure setting.

Definition 7.2. *We define the **causal parent probability** of $\boldsymbol{Y}$ given $\boldsymbol{X}$ as:*

$$\mathrm{P}(\boldsymbol{Y} || \mathit{par}(\boldsymbol{Y})) \triangleq \prod_t \mathrm{P}(Y_t | \mathit{par}(Y_t)). \tag{7.2}$$

Definition 7.3. *The **causal parent entropy** is formed from this probability distribution and is:*

$$H(\boldsymbol{Y} || \mathit{par}(\boldsymbol{Y})) \triangleq \mathrm{E}_{\mathrm{P}(\boldsymbol{X}, \boldsymbol{Y})}[-\log \mathrm{P}(\boldsymbol{Y} || \mathit{par}(\boldsymbol{Y}))]. \tag{7.3}$$

This entropy measure serves as the objective function that extends the maximum causal entropy approach to settings with additional imperfect information. The key distinction between this entropy measure and the previous causal entropy measure (Equation 4.13) introduced in Chapter 4 is that some of the variables in $\mathbf{X}$ may never be directly conditioned upon by any $\mathbf{Y}$ variable or may be conditioned on intermittently (along with past $\mathbf{Y}$ variables) over time in the causal parent entropy measure of this section.

Definition 7.4. *The **imperfect information maximum causal entropy optimization** maximizes the causal parent entropy (Equation 7.3) while matching expected and statistic-based feature func-*

tions, $\mathcal{F}(\boldsymbol{X}, \boldsymbol{Y}) \to \mathbb{R}^K$. It is formally defined as:

$$\max_{\{\mathrm{P}(Y_t|par(Y_t))\}} H(\boldsymbol{Y}||par(\boldsymbol{Y})) \tag{7.4}$$

$$\text{such that: } \mathrm{E}_{\mathrm{P}(\boldsymbol{X},\boldsymbol{Y})}\left[\sum_t \mathcal{F}_t(\boldsymbol{X}_{1:t}, \boldsymbol{Y}_{1:t})\right] = \mathrm{E}_{\tilde{P}(\boldsymbol{X},\boldsymbol{Y})}\left[\sum_t \mathcal{F}_t(X_t, Y_t)\right]$$

$$\forall_{t, Y_t, par(Y_t)} \ \mathrm{P}(Y_t|par(Y_t)) \geq 0$$

$$\forall_{t, par(Y_t)} \sum_{Y_t} \mathrm{P}(Y_t|par(Y_t)) = 1.$$

Just as in the general maximum causal entropy optimization (Equation 5.6), the convexity of the imperfect information optimization (Equation 7.4) when optimizing in terms of the conditional probability terms, $\{\mathrm{P}(Y_t|\mathrm{par}(Y_t))\}$, is not readily apparent. Indeed, only with further restrictions to the parent structure of variables do we establish the convexity of the optimization in Equation 7.4.

Definition 7.5. ***Perfect past decision recall*** *is a constraint on the parent sets of decision variables, $\boldsymbol{Y}$, that forces all past decision variables to be parents of future decision variables:*

$$\forall_t \ \boldsymbol{Y}_{1:t-1} \subseteq par(Y_t). \tag{7.5}$$

Convexity follows from the additional constraint of perfect decision recall (Definition 7.5).

Theorem 7.6. *The imperfect information maximum causal entropy optimization of Equation 7.4 constrained by perfect past decision recall can be formulated as a convex optimization problem.*

Proof (sketch). Consider the causal conditional probability distribution, $\{\mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T)\}$, of a particular conditioned variable, Y_t, its parent set, $\mathrm{par}(Y_t)$, and the variables that are **greater-ancestors** (*i.e.*, non-parent ancestors of Y_t):

$$\mathrm{granc}(Y_t) \triangleq \mathrm{anc}(Y_t) \cap \overline{\mathrm{par}(Y_t)}.$$

As before, future side information variables can have no causal influence over earlier conditioned variables; this is accomplished using the final constraints of Equation 5.6—marginalizing over non-ancestor variables. This constrains the causally conditioned probability distribution to factor into a product of ancestor-conditioned terms:

$$\mathrm{P}(\mathbf{Y}||\mathbf{X}) = \prod_t \mathrm{P}(Y_t|\mathrm{par}(Y_t), \mathrm{granc}(Y_t)).$$

Of the ancestor values in each conditional probability term: $\mathrm{P}(Y_t|\mathrm{par}(Y_t), \mathrm{granc}(Y_t))$, the granc set of variables are unobserved. Constraining these conditional probability distributions to be equivalent across varying greater-ancestor values,

$$\forall_{t, Y_t, \mathrm{par}(Y_t), \mathrm{granc}(Y_t)_1, \mathrm{granc}(Y_t)_2} \mathrm{P}(Y_t|\mathrm{par}(Y_t), \mathrm{granc}(Y_t)_1) = \mathrm{P}(Y_t|\mathrm{par}(Y_t), \mathrm{granc}(Y_t)_2),$$

enforces any imperfect information constraints that are implied by the variable dependence structure. When constrained by perfect past decision recall, this constraint is a linear function of causally conditioned probabilities. □

Remark 7.7. *Without perfect past decision recall, the imperfect information maximum causal entropy optimization is non-convex in general.*

This property of perfect decision recall denotes an important boundary for the applicability of the causal entropy approach; without it, convexity and strong duality, upon which the maximum entropy approach relies, do not hold (Remark 7.7). The inefficiency of reasoning in the general imperfect past decision setting should not be particularly surprising; it could be employed to represent Nash equilibria inference problems in a wide range of game settings that are known to be non-polynomial (assuming the complexity hierarchy does not collapse).

7.2.2 Distribution Form

Following the formulation of Theorem 7.6, we obtain the form of the distribution for the imperfect information setting with perfect past decision recall.

Theorem 7.8. *The maximum causal entropy probability distribution for the imperfect information setting with perfect past decision recall (Theorem 7.6) is distributed according to the following recurrence relationship:*

$$
\begin{aligned}
P_\theta(Y_t|\mathit{par}(Y_t)) &= \frac{Z_{Y_t|\mathit{par}(Y_t),\theta}}{Z_{\mathit{par}(Y_t),\theta}} \qquad (7.6)\\
\log Z_{\mathit{par}(Y_t),\theta} &= \log \sum_{Y_t} Z_{Y_t|\mathit{par}(Y_t),\theta} \\
&= \operatorname*{softmax}_{Y_t} \left(\mathrm{E}_{\mathrm{P}(\mathit{par}(Y_{t+1})|\mathit{par}(Y_t),Y_t)} \left[\log Z_{\mathit{par}(Y_t),\theta} | \mathit{par}(Y_t), Y_t \right] \right) \\
Z_{Y_t|\mathit{par}(Y_t),\theta} &= e^{\left(\mathrm{E}_{\mathrm{P}(\mathit{par}(Y_{t+1})|\mathit{par}(Y_t),Y_t)} \left[\log Z_{\mathit{par}(Y_t),\theta} | \mathit{par}(Y_t),Y_t \right] \right)} \\
Z_{\mathit{par}(Y_{T+1}),\theta} &= e^{\theta^\top \mathcal{F}(\boldsymbol{X},\boldsymbol{Y})},
\end{aligned}
$$

where the final set of parents for the "after last" Y variable is the complete set of variables: $\mathit{par}(Y_{T+1}) \triangleq \boldsymbol{X} \cup \boldsymbol{Y}$.

It is similar to the feature-matching maximum causal entropy distribution, except that additional variables can be latent and are then marginalized over given the parents of each Y_t.

7.2.3 Perfect Recall Reduction

One important special case of imperfect information that simplifies the distribution of Equation 7.6 is perfect recall, which limits uncertainty to side information variables that will be revealed at

some point in time (or are never revealed). This specifically excludes past decisions or influences on decisions that are later "forgotten."

Definition 7.9. *In the **perfect recall** sequential prediction setting, each conditioned variable, Y_t, is constrained to be conditioned on all previous conditioned variables and variables on which they depended. This constrains the parent sets as follows:*

$$\forall_t \, \textit{par}(Y_t) \cup Y_t \subseteq \textit{par}(Y_{t+1}). \tag{7.7}$$

Imperfect information modeling problems with perfect recall structure (Definition 7.9) can be reduced to the maximum causal entropy inverse optimal control setting of Chapter 6 via Theorem 7.10.

Theorem 7.10. *In the perfect recall setting, the imperfect information maximum causal entropy distribution of Equation 7.4 can be reduced to a non-latent maximum causal entropy model by employing expectations to obtain side information dynamics,*

$$\begin{aligned} &P'\left(\textit{par}(Y_t) | Y_{t-1}, \textit{par}(Y_{t-1})\right) = \\ &\quad \mathrm{E}_{\mathrm{P}(\boldsymbol{X}|Y_{t-1},\textit{par}(Y_{t-1}))}\left[\mathrm{P}(\textit{par}(Y_t)|\boldsymbol{X}, Y_{t-1}, \textit{par}(Y_{t-1}))\ \mathrm{P}(\boldsymbol{X}|Y_{t-1}, \textit{par}(Y_{t-1}))\middle| Y_{t-1}, \textit{par}(Y_{t-1})\right], \end{aligned}$$

and expected statistic-based features,

$$\mathcal{F}'_t\left(Y_t, \textit{par}(Y_t)\right) = \mathrm{E}_{\mathrm{P}(\boldsymbol{X}|\textit{par}(Y_t))}\left[\mathcal{F}_t(\boldsymbol{X}, \boldsymbol{Y}_{1:t-1}))\,\mathrm{P}(\boldsymbol{X}|\textit{par}(Y_t))\middle|\textit{par}(Y_t)\right].$$

Additionally, strict perfect recall is often not necessary; parent sets that satisfy perfect recall can often be reduced without affecting the corresponding maximum causal entropy probability distribution.

Definition 7.11. *A subset of conditioning variables, $\textit{par}_{\mathcal{C}}(Y_t) \subseteq \textit{par}(Y_t)$, (i.e., side information, $\boldsymbol{X}$, or previous conditioned variables, $\boldsymbol{Y}_{1:t-1}$,) is **relevant** in the maximum causal entropy setting if it is the minimum size set of variables such that:*

$$\mathrm{P}(Y_t|\textit{par}(Y_t)) = \mathrm{P}(Y_t|\textit{par}_{\mathcal{C}}(Y_t)),$$

where $\textit{par}_{\mathcal{C}}(Y_t)$ is the set of parents for Y_t that are members of set $\mathcal{C}$, and $\mathrm{P}(Y_t|\textit{par}_{\mathcal{C}}(Y_t))$ is the distribution obtained by maximizing the causal parent entropy (Equation 7.4) according to this alternate conditioning structure.

Any parents of a conditioned variable that are not relevant (Definition 7.11) can be removed from the parent set, reducing the complexity of the resulting maximum causal entropy model's parametrization.

7.3 Example Maximum Causal Entropy Influence Diagram Representations

Many sequential data modeling problems can be expressed as maximum causal entropy influence diagrams. This is a natural consequence of many optimal control problems being representable as influence diagrams (Section 2.2.3). We now review a few classes of decision prediction tasks to illustrate the breadth of applicability of the maximum causal entropy influence diagram.

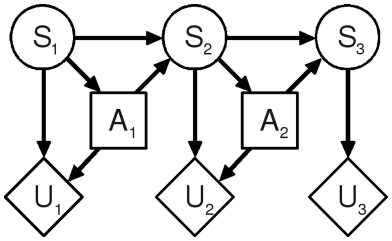

Figure 7.1: The maximum causal entropy influence diagram graphical representation for maximum causal entropy inverse optimal control. For Markov decision processes: $U_t(S_t, A_t) = \theta^\top \mathbf{f}_{S_t,A_t}$, and for linear quadratic regulation models: $U_t(\mathbf{s}_t, \mathbf{a}_t) = \mathbf{s}_t^\top \mathbf{Q}\mathbf{s}_t + \mathbf{a}_t^\top \mathbf{R}\mathbf{a}_t$. Absolute conditioning on $\{\mathbf{f}_{S,A}\}$ or $\mathbf{Q}$ and $\mathbf{R}$ is suppressed, as are perfect recall dependencies that are irrelevant.

The model of maximum causal entropy inverse optimal control (Chapter 6) is depicted in Figure 7.1. The Markov decision process and linear quadratic regulation settings can each be represented with different forms of utility node functions. Note that perfect past decision recall edges are not present. However, descendants' utility nodes are conditionally independent of past decisions given parents of an action. Thus, past decisions are irrelevant and the perfect past decision recall requirement is not violated.

The fully observable control prediction task of Figure 7.1 is extended to the setting where the state is only partially-observed through observation variables in Figure 7.2. This model satisfies the perfect recall constraint since all influences on past decisions and decisions are parents of future decision nodes.

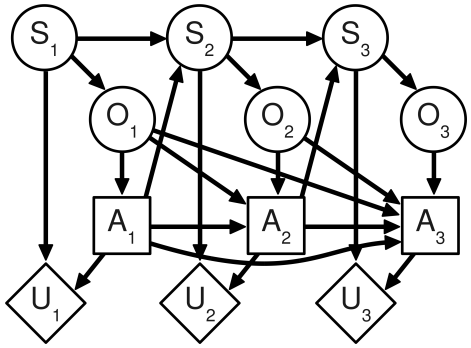

Figure 7.2: The maximum causal entropy influence diagram graphical representation for maximum causal entropy inverse optimal control in a partially observable system. The state variables are only partially observed via observation variables, $O_{1:T}$, that are distributed according to a known conditional distribution, $\mathrm{P}(O_t|S_t)$.

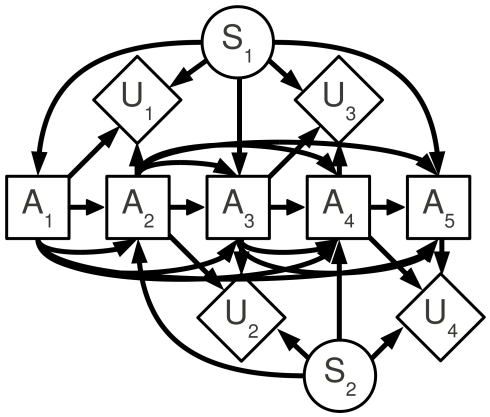

Figure 7.3: The extensive-form game setting where players have access to private information, S_1 and S_2, and take sequential decisions. Recall of all past actions is provided by the sets of edges connecting all decisions.

Figure 7.3 shows a maximum causal entropy influence diagram for an extensive-form game where each player has private information that is only conveyed to the other players through a choice of actions. The decision problem of this influence diagram satisfies the perfect past decision recall property.

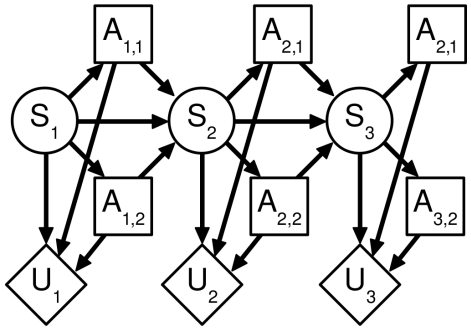

Figure 7.4: The Markov game setting where the state of the game changes according to known Markovian stochastic dynamics, $\mathrm{P}(S_{t+1}|S_t, A_{t,1}, A_{t,2})$, and the players share a common reward.

Figure 7.4 shows a maximum causal entropy influence diagram for a Markov game setting where two players act simultaneously without knowing the other player's action. This setting does not satisfy the perfect decision recall property because $A_{t,1}$ and $A_{t,2}$ have no recall relationship. We investigate settings where the actions at each timestep are correlated in Chapter 8.

7.4 Discussion

In this chapter, we have expanded the maximum causal entropy approach for matching empirical statistics of data to settings where in addition to future latent variables, past and current variables can be latent also.

The four imperfect information settings are covered to various degrees by the maximum causal entropy formulations introduced thus far, as shown in Table 7.2.

Table 7.2: Coverage of the four imperfect information settings by different maximum causal entropy variants.

	Latent future information	Partial observability	Latent past conditioning	Latent past decisions
Inverse optimal control maximum causal entropy	✓			
Perfect recall maximum causal entropy	✓	✓		
Perfect past decision recall maximum causal entropy	✓	✓	✓	
Imperfect recall maximum causal entropy	✓	✓	✓	✓

The perfect recall maximum causal entropy formulation (Definition 7.9) has the advantage of reducing to the inverse optimal control model, making inference relatively straight-forward. The perfect past decision recall formulation extends the applicability to settings where side information on which past decisions were based is latent. Finally, the imperfect information maximum causal entropy formulation (Definition 7.4) supports all of these forms of imperfect information, but is no longer a convex optimization problem.

The problem of recovering explanatory utility functions for influence diagrams was previously investigated for inconsistent behavior by assuming for each instance of behavior, a noisy copy of the true utility function was sampled and then optimal behavior under that noisy utility function employed to obtain the demonstrated behavior sequence (Nielsen & Jensen, 2004). Markov chain Monte Carlo techniques are then employed to learn the mean and variance of the distribution from which noisy utility functions are drawn. We view the maximum causal entropy influence diagram a significant improvement over this inefficient learning procedure since the convexity of its objective function enables much more efficient and more exact learning.

Chapter 8

Strategic Decision Prediction via Maximum Causal Entropy

"Maybe all one can do is hope to end up with the right regrets."
— Arthur Miller (Playwright, 1915–2005).

In this chapter, we investigate settings where multiple *agents* or *players* take actions within sequential stochastic games. Equilibria solution concepts, such as Nash equilibria (NE) (Nash, 1951) and correlated equilibria (CE) (Aumann, 1974) are important constructs for these games that provide certain individual or group performance guarantees. However, from a machine learning perspective, existing equilibria concepts are often not useful for prediction, because they do not fully specify a unique strategy profile. Instead they may specify a polytope or point set of possible expected utility outcome vectors, making strategy prediction under-specified in general without additional assumptions. We introduce the **maximum causal entropy correlated equilibria** solution concept for predictive purposes based on the principle of maximum causal entropy with rationality constraints that prevent any regret from changing actions from that prescribed by the joint distribution in this chapter.

8.1 Game Theory Background

8.1.1 Classes of Games

The canonical type of game studied within game theory is the one-shot game with a matrix of payoffs for the players (Definition 8.1).

Definition 8.1. *A **normal-form game**, $\mathcal{G}_N$, is defined by a set of players $i = 1...N$, a set of actions for each player, $a_{i,j} \in A_i$, and a utility vector $U_{a_{1,j_1},...,a_{N,j_N}} \in \mathbb{R}^N$, specifying the* payoffs *to each player for every combination of actions.*

In a normal-form game, each player ($i = 1...N$) selects an action (a_i) simultaneously without knowledge of other players' selected actions. A utility vector for each combination of actions, $U_{a_{1,j_1},...,a_{N,j_N}} \in \mathcal{R}^N$, specifies a numerical **payoff**, $U_{\mathbf{a},i}$, to each player (i).

Table 8.1: The prisoner's dilemma normal-form game. Two prisoners jointly receive the minimal sentence if they both remain silent, but each has an incentive to (unilaterally) confess.

	Silence	Confess
Silence	0.5 years, 0.5 years	10 years, 0 years
Confess	0 years, 10 years	5 years, 5 years

The prisoner's dilemma (Table 8.1) is a classic example of a normal form game.

Definition 8.2. *A **Markov game** (Filar et al., 1997) (also known as a **stochastic game**) is defined by a set of states (S) representing the joint states of N players, a set of individual actions (A) for each player, a probabilistic state transition function, $T : S \times A_{1:N} \rightarrow \mathrm{P}(S')$ specifying the distribution for next state, S', and a utility function, Utility$_i$: $S \times A_{1:N} \rightarrow \mathbb{R}$ for each player $i = 1 : N$.*

Markov games (Definition 8.2) generalize normal-form games to sequential action settings. They also generalize Markov decision problems by incorporating multiple players who act based upon their individual utility functions and jointly influence the future state of the game with their actions, as shown in Figure 8.1.

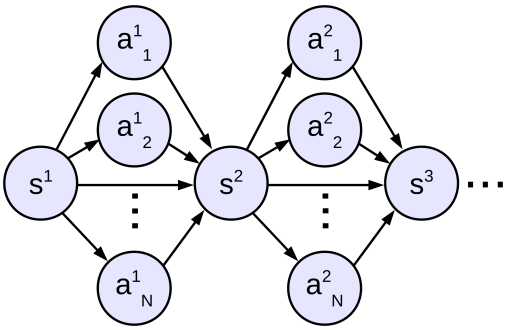

Figure 8.1: The sequence of states and (Markovian) actions of a Markov game. Actions at each time step can either be correlated (*i.e.*, dependently distributed based on past actions and states or an external signaling device), or independent.

Definition 8.3. *An **extensive-form game** is defined by a tree of action sequences and state outcomes where each agent is incapable of distinguishing between multiple states in the tree.*

Extensive form games (Definition 8.3) further extend the Markov game setting in a way where unique sequences of states and actions may end up in a meta-state that is indistinguishable by the agent from other states. This allows situations with incomplete information, such as imperfect recall, to be represented within the model.

Game players choose a **strategy profile**, π, specifying next actions for each situation that are either **mixed** (stochastic) or **pure** (deterministic); and either **correlated** (joint functions) or **independent** (factor by players) based on a (discounted, $0 < \gamma \leq 1$) cumulative expected utility:

$$\text{ExpectUtil}_i^\pi(a_{1:N}^t, s^t) \triangleq \mathrm{E}_{\pi(\mathbf{s}^{t+1:T}, \mathbf{a}_{1:N}^{t+1:T} | a_{1:N}^t, s^t)} \left[\sum_{\tau \geq t} \gamma^\tau \text{Utility}_i(s^\tau, a_{1:N}^\tau) \middle| a_{1:N}^t, s^t, \pi \right] \tag{8.1}$$

for player i under the joint strategy profile, π. We assume in Equation 8.1 and throughout this chapter that the strategy profile is mixed and **Markovian**[1], meaning it depends only on the current state and timestep. In a **stationary** mixed strategy profile, the joint action probabilities for each state are the same for all time steps.

To obtain strategy profiles, it is useful to consider the amount of utility gained when taking a deviation action, $a_i^{t\prime}$, instead of a provided action, a_i^t, when: all players' actions except player i's, denoted a_{-i}^t, are known (Equation 8.2); or when other players' actions, a_{-i}^t, are unknown and averaged over according to the strategy profile (Equation 8.3):

$$\text{ExpectDevGain}_i^\pi(a_{1:N}^t, s^t, a_i^{t\prime}) \triangleq \text{ExpectUtil}_i^\pi(\{a_{-i}^t, a_i^{t\prime}\}, s^t) - \text{ExpectUtil}_i^\pi(a_{1:N}^t, s^t) \tag{8.2}$$

$$\text{ExpectRegret}_i^\pi(a_i^t, a_i^{t\prime}, s^t) \triangleq \mathrm{E}_{P_\pi(a_{-i}^t | a^t, s^t)} \Big[\text{ExpectDevGain}_i^\pi(a_{1:N}^t, s^t, a_i^{t\prime}) \Big| a_i^t, s^t \Big]. \tag{8.3}$$

These quantities are useful for assessing the rationality of multi-player strategies.

8.1.2 Equilibria Solution Concepts

Defining equilibria solution concepts and finding corresponding equilibria in multi-agent settings are important problems for applications ranging from conflict resolution to market-making. Perhaps the most widely-known solution concept in game theory is the Nash equilibrium.

Definition 8.4. *A **Nash equilibrium** for a game is defined as a fixed point where no agent has positive deviation regret and agents' actions are independent.*

Though any finite game is guaranteed to have a mixed strategy Nash equilibrium (Nash, 1951), only exponential-time NE-finding algorithms are known to exist (Lemke & Howson Jr, 1964) for general-sum normal-form games, and associated decision problems regarding the characteristics of possible NE are NP-hard (Gilboa & Zemel, 1989).

[1] Markovian equilibria strategy profiles are a consequence of the MCECE formulation and a commonly assumed constraint imposed on equilibria in other solution concept formulations.

When extending equilibria from the normal-form setting to the dynamic setting of Markov games, the concept of **sub-game equilibria**—that even in states that are not reached under a policy, the expected utilities of those states are based on the equilibrium of a game starting in that state—is an important requirement.

Definition 8.5. *A **correlated equilibrium** (CE) (Aumann, 1974) for a Markov game is a mixed joint strategy profile, π^{CE}, where no expected utility gain is obtained for any player by substituting an action, $a_i^{t'}$ that deviates from the strategy. This is guaranteed with the following set of constraints:*

$$\forall_{t,i,s^t,a_i^t,a_i^{t'}} \textit{ExpectRegret}_i^{\pi^{CE}}(a_i^t, a_i^{t'}, s^t) \leq 0. \tag{8.4}$$

Under this set of constraints, given the agent's prescribed action, a_i^t, there is no benefit for deviating (under the distribution of other players' actions, $P(a_{-i}^t|a_i^t)$). We further require that regrets be defined according to sub-game correlated equilibria.

Unlike Nash equilibria (Nash, 1951), which require independent player strategies, players in a CE can coordinate their actions to obtain a wider range of expected utilities. Traffic lights are a canonical example of a **signaling device** designed to produce CE strategies. However, external signaling mechanisms are not necessarily required—players can coordinate their actions from their history of past actions in sequential game settings. The deviation regret constraints (Equation 8.4) define a convex polytope of CE solutions in the N-dimensional space of players' joint utility payoffs (Figure 8.2).

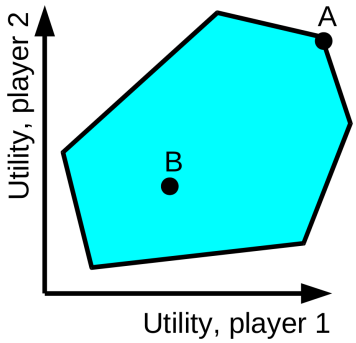

Figure 8.2: A correlated equilibria polytope with a correlated-Q equilibrium (Definition 8.6) payoff at point A that maximizes the average utility and a maximum entropy correlated equilibrium at point B (Definition 8.9) that provides predictive guarantees.

Consider the single shot, two-player prisoner's dilemma of Table 8.1. For each of the two players and two actions, there is one alternate action. Thus there are a total of four linear constraints for the prisoner's dilemma setting (and more generally, any two-player, two-action normal-form game).

In general, for a single-shot (*i.e.*, normal-form) game, there are $O(N|A|^2)$ regret constraints that are linear in a total of $O(|A|^N)$ strategy variables, $\{\pi(a_{1:N})\}$, and CE solutions can be efficiently obtained (*i.e.*, in the same polynomial time as the size of the Utility matrix specification) by solving a linear program (LP) or a convex program (CP):

$$\max_{\pi} f_0(\pi(A_{1:N})) \tag{8.5}$$
$$\text{such that: } \forall_{i,a_i,a_{i'}} \sum_{a_{-i}} \pi(a_{1:N}) \big(\text{Utility}\left(\{a_{-i}, a_i{}'\}\right) - \text{Utility}(a_{1:N})\big) \leq 0,$$
$$\forall_{a_{1:N}} \pi(a_{1:N}) \geq 0, \text{ and } \sum_{a_{1:N}} \pi(a_{1:N}) = 1.$$

depending on whether the objective function, f_0, is linear or (negative) convex.

Definition 8.6. *A **correlated-Q equilibria** (CE-Q) (Greenwald & Hall, 2003) employs a linear or convex function of strategy probabilities for the selection metric objective of Equation 8.5 to obtain utility-unique strategy profiles*[2].

A number of objectives for CE-Q have been proposed (Greenwald & Hall, 2003):

- **Utilitarian** (*u*CE-Q) maximizes the sum of players' utilities, $\sum_{i=1}^{N} \mathrm{E}_\pi[\text{Utility}_i(a_{1:N})]$;
- **Dictatorial** (*d*CE-Q) maximizes a specific player's utility, $\mathrm{E}_\pi[\text{Utility}_i(a_{1:N})]$;
- **Republican** (*r*CE-Q) maximizes the highest player's utility, $\max_i \mathrm{E}_\pi[\text{Utility}_i(a_{1:N})]$; and
- **Egalitarian** (*e*CE-Q) maximizes lowest player's utility, $\min_i \mathrm{E}_\pi[\text{Utility}_i(a_{1:N})]$.

These strong assumptions about players' preferences constrain CE-Q solutions to reside on the hull of the most-positive quadrant of the CE polytope (Northeast in Figure 8.2). They do not necessarily specify a unique strategy profile or joint set of payoffs. However, they do specifically require that a corner of the joint utility polytope is covered by a CE-Q solution. More generally, strategies from the other polytope quadrants are also possible. Various forms of punishment, for example, such as grim-trigger strategies, have been recognized as viable sub-game strategies that disincentivize a player's undesirable actions. We augment these existing CE-Q with a few based on punishment:

- **Disciplinarian** (*x*CE-Q) minimizes a specific player's utility, $E_\pi[\text{Utility}_i(a_{1:N})]$; and
- **Inegalitarian** (*i*CE-Q) maximizes utility differences between two (groups of) different players, $E_\pi[\text{Utility}_i(a_{1:N}) - \text{Utility}_j(a_{1:N})]$.

[2]Unique strategy profiles are not guaranteed by the CE-Q solution concept—multiple actions can provide the same vector of player utilities, and the CE-Q strategies may comprise an entire facet of the CE polytope. We ignore this ambiguity and employ a single CE-Q from the set of CE-Q strategy profiles in this work.

Definition 8.7. *The **maximum entropy correlated equilibria** (MaxEntCE) solution concept for normal-form games (Ortiz et al., 2007) selects the unique CE with the* fewest *additional assumptions (Figure 8.2) by employing Shannon's information entropy as the objective function,* $-\sum_{a_{1:N}} \pi(a_{1:N}) \log \pi(a_{1:N})$*, as advocated by the **principle of maximum entropy** (Jaynes, 1957).*

Since a Nash equilibria is guaranteed to exist (Nash, 1951) and a Nash equilibria is a special case of correlated equilibria that is additionally constrained to have independent actions: $\mathrm{P}(a_{1:K}) = \prod_k \mathrm{P}(a_k)$, a correlated equilibria is always similarly guaranteed to exist. Of those existing, one must maximize the entropy measure. However, the relaxation of the non-linear independent action constraint that differentiates Nash equilibria from correlated equilibria enables algorithms to efficiently obtain correlated equilibria (Papadimitriou & Roughgarden, 2005) and maximum entropy correlated equilibria (Ortiz et al., 2007) for a wider class of problems than can be addressed using any algorithms for finding Nash equilibria (assuming the complexity hierarchy does not collapse).

Table 8.2: The game of Chicken and its correlated equilibria strategy profiles.

	Stay	Swerve
Stay	0,0	4,1
Swerve	1,4	3, 3

CE 1

0	1
0	0

CE 2

0	0
1	0

CE 3

0	$\frac{1}{3}$
$\frac{1}{3}$	$\frac{1}{3}$

CE 4

$\frac{1}{4}$	$\frac{1}{4}$
$\frac{1}{4}$	$\frac{1}{4}$

Consider the game of Chicken (where each player hopes the other will *Swerve*) and the correlated equilibria that define its utility polytope in Table 8.2. *CE 1* and *CE 2* are both dictatorial and inegalitarian CE (for different players) and republican CE (but ambiguous). *CE 3* is a utilitarian CE and an egalitarian CE. *CE 4* is the maximum entropy CE. Its predictive guarantee is apparent: all other CE have infinite log-loss for at least one other CE; the MaxEntCE is the only CE that assigns positive probability to the $\{$Stay, Stay$\}$ action combination. We extend these predictive guarantees to the Markov game setting in this work.

8.2 Maximum Causal Entropy Correlated Equilibria

We are motivated by two problems within the Markov game setting (Filar et al., 1997) characterized by the dynamic interaction of multiple agents with a stochastic environment. The first is where a trusted third party (or signaling mechanism) must coordinate the behavior of agents in a way that is both:

- A correlated equilibria that can be found efficiently; and
- Sensitive to the amount of information revealed about the agents' underlying motives.

The second is the problem of predicting the behavior of agents that are assumed to be acting according to an unknown correlated equilibria, which can naturally be arrived at without explicit coordination by employing certain strategies in repeated games (Hart & Mas-Colell, 2000).

8.2.1 Formulation

Extension of the MaxEntCE solution concept (Ortiz et al., 2007) to the Markov game setting is not straight-forward. The first difficulty is that the deviation regret constraints of normal-form games (Equation 8.5), which are linear in the unknown mixed strategy variables, contain expectations over future actions (Equation 8.4) when extended to the Markov game setting, creating non-linear constraints that are products of the unknown variables.

Theorem 8.8. *A linear program/convex program formulation of CE for Markov games is possible by considering as variables the entire sequence of joint player actions for the sequence of revealed states, $\eta(A_{1:N}^{1:T}|S^{1:T})$, and employing appropriate inequality constraints (deviation regret guarantees) and equality constraints (forcing the strategy over sequences to factor into products of Markovian strategies) on marginal distributions using linear functions of $\eta(A_{1:N}^{1:T}|S^{1:T})$ variables.*

Naïvely formulating the Markov game CE strategy profiles into an LP/CP is possible (Theorem 8.8), but the number of constraints and variables grows exponentially with the time horizon[3]. The second difficulty is that there are many entropy measures based on joint, conditional, and marginal probability distributions that could be applied as objective functions. For example, the **joint entropy** is a natural entropy measure to consider. It is a combination of the entropy of the strategy profiles' actions and the entropy of the state dynamics:

$$H(a_{1:N}^{1:T}, s^{1:T}) = \sum_{t=1}^{T}(H(a_{1:N}^{t}|s^{1:t}, a_{1:N}^{1:t-1}) + H(s^{t+1}|s^{1:t}, a_{1:N}^{1:t})) = H(a_{1:N}^{T}||s^{T}) + H(s^{T}||a_{1:N}^{T-1}). \tag{8.6}$$

Since the transition dynamics are already known in this problem setting, the uncertainty associated with those dynamics is irrelevant. Maximizing the joint entropy generally makes the players' strategies *less* uncertain than otherwise possible by adding the assumption that players care about the uncertainty that actions supply beyond expectations over their random outcomes. The reason for this extends directly from the decision theory setting illustrated by Figure 6.3 in Section 6.3.2.

We instead advocate the **causally conditioned entropy** measure (Kramer, 1998) $H(\mathbf{A}_{1:N}^{T}||\mathbf{S}^{T}) \triangleq \sum_t H(a_{1:N}^{t}|a_{1:N}^{1:t-1}, s^{1:t})$, employed throughout this thesis. For the possible sequences of states and actions through a Markov game, it corresponds to the uncertainty associated with only the actions in such sequences. It is based on the **causally conditioned probability distribution**,

[3]The number of constraints is also exponential in the number of players. However, this is unavoidable in general game settings since the size of the payoff matrix grows exponentially with the number of players. We are thus only concerned with reducing the exponential dependence on the time horizon in this work.

$\mathrm{P}(\mathbf{A}^T_{1:N}||\mathbf{S}^T) \triangleq \prod_t \mathrm{P}(a^t_{1:N}|a^{1:t-1}_{1:N}, s^{1:t})$, which conditions each set of correlated actions only on actions and states that have been revealed at that point in time and not on future states, as in the conditional probability distribution $\mathrm{P}(\mathbf{A}_{1:N}|\mathbf{S}) = \prod_t \mathrm{P}(a^t_{1:N}|a^{1:t-1}_{1:N}, s^{1:t}, s^{t+1:T})$.

Definition 8.9. *A **maximum causal entropy correlated equilibrium** (MCECE) solution maximizes the causal entropy while being constrained to have no action deviation regrets*[4]*:*

$$\pi^{MCECE} \triangleq \underset{\pi}{\operatorname{argmax}}\, H(a^T_{1:N}||s^T) = \underset{\pi}{\operatorname{argmax}}\, \mathrm{E}_{a^{1:T}, s^{1:T}}\left[\sum_{t=1}^{T} -\log \mathrm{P}(a^t_{1:N}|s^t)\right] \tag{8.7}$$

$$\text{such that: } \forall_{t,\, i,\, a^t_i,\, a^{t'}_i,\, s^t}\ \textit{ExpectRegret}^{\pi}_i(a^t_i,\ a^{t'}_i,\ s^t) \leq 0$$

$$\forall_{t,\, s^t,\, a^t_{1:N}}\ \mathrm{P}(a^t_{1:N}|s^t) \geq 0, \quad \forall_{t,\, s^t} \sum_{a^t_{1:N}} \mathrm{P}(a^t_{1:N}|s^t) = 1,$$

$$\pi \textit{ factors as: } \mathrm{P}(\boldsymbol{A}^T_{1:N}||\boldsymbol{S}^T), \textit{ and given: } \{\mathrm{P}(s^{t+1}|s^t, a^t)\}.$$

We further constrain the strategy profile to have sub-game equilibria, meaning that even in states that are unreachable under the strategy profile and state dynamics, the strategy profile is constrained to satisfy Equation 8.7 in all sub-games starting from those states.

8.2.2 Properties

Based on the view of conditional entropy as a measure of predictability (Cover & Thomas, 2006), the MCECE solution concept extends two important predictive guarantees to the multi-agent setting:

Theorem 8.10 (extension of (Ortiz et al., 2007))**.** *Given an MCECE strategy profile, no player may decrease the predictability of her action sequence without creating deviation regret for herself.*

Theorem 8.11 (extension of (Grünwald & Dawid, 2003))**.** *The MCECE solution strategy profile, π^{MCECE} minimizes the worst-case log prediction loss for the sequences of joint actions, i.e.,*

$$\inf_{\mathrm{P}(\boldsymbol{A}^T||\boldsymbol{S}^T)} \sup_{\tilde{P}(\boldsymbol{A}^T||\boldsymbol{S}^T)} \sum_{\boldsymbol{A},\boldsymbol{S}} \tilde{P}(\boldsymbol{A}, \boldsymbol{S}) \log \mathrm{P}(\boldsymbol{A}^T||\boldsymbol{S}^T), \tag{8.8}$$

of all the CE-satisfying deviation regret constraints, where $\tilde{P}(\boldsymbol{A}^T||\boldsymbol{S}^T)$ is the (worst possible for prediction) empirical CE strategy and the joint, $\tilde{P}(\boldsymbol{A}, \boldsymbol{S})$, is the distribution of states and actions under that strategy profile and the known state transition dynamics.

[4]We employ a Markovian formulation for simplicity of presentation. Note that it arises without additional assumptions if a history-dependent strategy profile, $\mathrm{P}(a^t_{1:N}|s_{1:t}, a^{1:t-1}_{1:N})$, is employed, since the MCECE strategy profile ultimately reduces to a Markovian strategy profile given the standard Markovian dynamics and Markovian payoffs assumptions of Markov games.

Theorem 8.11 is particularly relevant to our machine learning perspective, because it justifies the MCECE strategy profile as a robust predictive model. It also provides a natural interpretation in gambling settings.

Remark 8.12. *Given that players of a stochastic game behave according to some unknown correlated equilibrium, betting according to the MCECE distribution maximizes the worst-case growth rate of winnings in this setting.*

Alternate entropy measures, such as the joint entropy, provide guarantees that are mathematically similar to Theorem 8.10 and Theorem 8.11 but that do not match the Markov game setting.

Corollary 8.13 (of Theorem 8.8)**.** *The solution for the MCECE optimization (Equation 8.7) can be obtained as the result of a convex program.*

Despite being obtainable by convex optimization (Corollary 8.13), the straight-forward convex program is not practical due to its exponential growth in T. We present an efficient computational formulation in Section 9.3.

8.2.3 Distribution Form

We obtain the general form of the MCECE distribution from its optimization. Despite the non-compact, history-dependent formulation of the naïve MCECE convex program, the strategy profile can be expressed compactly for Markov games.

Lemma 8.14. *The MCECE strategy profile for a Markov game is also Markovian.*

Theorem 8.15. *The MCECE strategy profile, $\pi_\lambda^{MCECE}(a_{1:N}^t|s^t)$, has the following recursive form (with $\lambda \geq 0$):*

$$\pi_\lambda^{MCECE}(a_{1:N}^t|s^t) \propto e^{-\left(\sum_{i,a_i^{t\prime}} \lambda_{i,s^t,a_i^t,a_i^{t\prime}} \mathit{ExpectDevGain}_i^{\pi^{MCECE}}\left(a_{1:N}^t,s^t,a_i^{t\prime}\right)\right)+\mathit{ExpectEnt}\left(a_{1:N}^t,s^t\right)}, \tag{8.9}$$

where $\mathit{ExpectEnt}(a_{1:N}^t, s^t) \triangleq \mathrm{E}_{a_{1:N}^{t+1},s^{t+1}}\left[\mathit{ExpectEnt}(a_{1:N}^{t+1}, s^{t+1}) + H(a_{1:N}^{t+1}|s^{t+1})\middle|a_{1:N}^t, s^t\right]$.

We discuss algorithms for recursively generating this distribution in Section 9.3.1 and for obtaining λ parameters in a sequence of sub-problems for this recursion in Section 10.1.3.

8.3 Discussion

In this chapter, we have applied the maximum causal entropy approach to multi-player sequential game settings. The result is a predictive model of correlated equilibrium strategy profiles that provides worst-case log-loss predictive guarantees in sequential game settings. This application demonstrates the versatility of the maximum causal entropy approach to settings beyond standard

decision theory. It also illustrates the use of inequality constraints to enforce assumed rationality characteristics of jointly rational behavior.

Dudık & Gordon (2009) previously applied maximum entropy to model behavior in extensive-form games. In that formulation, nature is treated as an additional player and the entropy associated with nature's actions is also maximized. As we argued in Section 6.3.2, when the dynamics of the game are assumed to be common knowledge, maximizing this entropy introduces sensitivity to the transition dynamics and does not provide predictive guarantees of players' strategies as a result. Additionally, while extensive-form games generalize the classes of games we consider in this chapter, strong duality does not hold; instead, only locally optimal model parameters trained from data can be guaranteed.

8.3.1 Combining Behavioral and Rationality Constraints

The approach of this chapter is quite agnostic; no observed strategies are employed to predict future behavior. Given additional knowledge of past behavior, the employed rationality constraints (inequalities) can be augmented with behavioral constraints (equalities) that force the predictive distribution to match demonstrated behavior strategies (in expectation) as in the inverse optimal control setting of Chapter 6, and the maximum causal entropy influence diagram setting of Chapter 7.

8.3.2 Infinite-horizon Games

Extending from finite-horizon to infinite-horizon games is not as straight-forward in the multi-player settings as it is in single-agent decision theory. Namely, asymmetric turn-taking behavior arises as rational behavior in many games, and any finite-behavior game is then very sensitive to the specific termination conditions of the game. Thus, approximating the infinite-horizon game with a very large finite-horizon game does not necessarily converge (Zinkevich et al., 2006). Murray & Gordon (2007) and MacDermed & Isbell (2009) attempt to efficiently extend value-iteration to propagate the entire polytope of correlated equilibria dynamically. Using approximations, this can be made efficient in practice. This approach holds promise for the maximum causal entropy approach, but we do not investigate it in this thesis.

Part III

Algorithms

Overview of Part III

Part III of this thesis leverages the theory of maximum causal entropy developed in Part II to formulate efficient algorithms for reasoning and training within the maximum causal entropy models of behavior.

Chapter 9 presents algorithms for efficient inferences of policies within the Markov decision process, linear-quadratic regulation, and influence diagram frameworks and strategies within the Markov game framework. Many of these algorithms are "softened" variants of the Bellman equation or similar sequential dynamic programming algorithms.

Chapter 10 presents algorithms for training maximum causal entropy models from demonstrated behavior (or, in the case of multi-player game settings, simply to enforce rationality requirements). These algorithms are simple gradient-based optimization techniques that rely on the convexity properties of the principle of maximum causal entropy from Part II to provide convergent optimality guarantees.

Chapter 11 investigates the setting where observed behavior is chosen in pursuit of some goal known to the actor but unknown to the observer. Bayesian inference methods are developed for inferring the latent goal with improved efficiency techniques for the non-adaptive (*i.e.*, deterministic dynamics) special case.

The algorithms of these three chapters together enable the maximum causal entropy approach to be efficiently applied to large-scale behavior modeling and prediction tasks investigated in Part IV of this thesis.

Chapter 9

Probabilistic Inference

"Computers are useless. They can only give you answers."
— Pablo Picasso (Artist, 1881–1973).

We begin our exploration of algorithms for the maximum causal entropy approach by focusing our attention on the development and analysis of algorithms for the task of efficient inference given model parameters within the maximum causal entropy models for feature expectation matching (Chapter 6), maximum causal entropy influence diagrams (Chapter 7), and predicting multi-agent actions (Chapter 8). Since approaches to parameter optimization (*i.e.*, learning) within maximum causal entropy models requires repeated inference for many different parameter values (Chapter 10), efficient inference is especially important for the maximum causal entropy approach to be applicable in large modeling domains.

9.1 Statistic-Matching Inference

We first consider inference algorithms for the statistic-matching setting of Chapter 6. As established by Theorem 6.8, there is a close connection between maximum causal entropy inference and stochastic optimal control based on a softened interpretation of the Bellman equation (Bellman, 1957). Our algorithms are developed along this line of thought, with some important specializations for the deterministic state transition dynamics setting, and significant differences in convergence characteristics from optimal control algorithms.

9.1.1 Policy and Visitation Expectations

We are interested in two closely-related inference tasks:

- Obtaining the maximum causal entropy policy, $\pi(A|S)$, for predictive purposes; and

- Obtaining the expected number of times a certain state or action will be executed under that policy, denoted D_{s_x} or $D_{a_{x,y}}$ for state x and action y in state x, respectively.

The latter quantities are needed to compute the statistics for model optimization purposes (*e.g.*, fitting to data).

From the soft Bellman equation interpretation of the maximum causal entropy distribution (Theorem 6.8), the policy is distributed according to:

$$\pi(a|s) = e^{Q^{\text{soft}}(s,a) - V^{\text{soft}}(s)}, . \tag{9.1}$$

The **state-action log partition function** , $Q^{\text{soft}}(s, a)$, and the **state log partition function**, $V^{\text{soft}}(s)$, of Equation 9.1 are related by the following recurrence:

$$\begin{aligned} Q^{\text{soft}}(a_t, s_t) &\triangleq \log Z_{a_t|s_t} \\ &= \mathrm{E}_{\mathrm{P}(s_{t+1}|s_t,a_t)}[V^{\text{soft}}(s_{t+1})|s_t, a_t] + \text{reward}(s_t, a_t) \qquad (9.2) \\ V^{\text{soft}}(s_t) &\triangleq \log Z_{s_t} \\ &= \operatorname*{softmax}_{a_t} Q^{\text{soft}}(a_t, s_t), \qquad (9.3) \end{aligned}$$

where $\operatorname{softmax}_x f(x) \triangleq \log \sum_x e^{f(x)}$.

We employ Algorithm 9.1 to dynamically compute the recurrence of Equation 9.2 and Equation 9.3. A terminal state reward function, $\phi(s)$, influences or constrains the terminal state distribution of state-action sequences. For example, if $\phi(s)$ is negative infinity except for one particular state, state-action sequences are then constrained to terminate in that state. The other goal settings (Section 6.1.5) can be similarly represented with appropriate choices of terminal state potential function, $\phi(s)$.

For T iterations of the softmax propagation of Algorithm 9.1 (the loop beginning at line 4), the algorithm's total running time is $O(T|A||S|)$ for the case where each action has a non-zero probability of transitioning to each state. We will usually be concerned with settings where the number of possible next states for each action is bounded by a small constant. In that case, the total running time reduces to $O(T(|\mathcal{A}| + |\mathcal{S}|))$.

We make frequent use of Algorithm 9.2 when performing exponentiated addition in the log space. It avoids underflow and overflow when computing log partition functions and other related quantities.

The policy is then recovered from the state log partition function using Equation 9.1 and Equation 9.2. We note that $\pi(a_{x,y}|s_x)$ does not necessarily normalize over actions, *i.e.*, often:

$$\sum_{a_{x,y}} \pi(a_{x,y}|s_x) < 1.$$

The remaining probability mass in any state that is not assigned to an action can be interpreted as an action that terminates the state-action sequence, which is distributed according to $e^{\phi(s) - V^{\text{soft}}(s)}$.

Algorithm 9.1 State log partition function calculation

Require: MDP, $\mathcal{M}_{\text{MDP}}$, and terminal state reward function, $\phi(s) \rightarrow \mathbb{R}$.
Ensure: state log partition functions, $V^{\text{soft}}(s_x)$.

for all States s_x **do**
  $V^{\text{soft}}(s_x) \leftarrow -\infty$
end for
while not converged **do**
  for all States s_x **do**
    $V^{\text{soft}}(s_x)' \leftarrow \phi(s_x)$
  end for
  for all Actions $a_{x,y}$ **do**
    $V^{\text{soft}}(s_x)' \leftarrow \text{softmax}\left(V^{\text{soft}}(s_x)', \text{reward}_{\mathcal{M}_{\text{MDP}}}(s_x, a_{x,y}) + \sum_z \text{P}(s_z|s_x, a_{x,y}) V^{\text{soft}}(s_z)\right)$
  end for
  for all States s_x **do**
    $V^{\text{soft}}(s_x) \leftarrow V^{\text{soft}}(s_x)'$
  end for
end while

Algorithm 9.2 Soft-maximum calculation

Require: inputs x_1, x_2
Ensure: $\text{softmax}(x_1, x_2) = \log\left(e^{x_1} + e^{x_2}\right)$

$\text{max}_{\mathbf{x}} \leftarrow \max(x_1, x_2)$
$\text{min}_{\mathbf{x}} \leftarrow \min(x_1, x_2)$
$\text{softmax}(x_1, x_2) \leftarrow \text{max}_{\mathbf{x}} + \log\left(1 + e^{\text{min}_{\mathbf{x}} - \text{max}_{\mathbf{x}}}\right)$

We then arrive at a forward-backward inference algorithm that can be extended to approximate the infinite time horizon.

From the policy action probabilities, a "forward pass" algorithm is employed to calculate visitation frequencies for states and actions, as shown by Algorithm 9.3. Action visitation frequencies are easily recovered from state visitation frequencies and the stochastic policy, since: $D_{a_{x,y}} = D_{s_x} \pi(a_{x,y}|s_x)$.

Similarly to Algorithm 9.1, the "forward" inference of Algorithm 9.3 is $O(T|\mathcal{A}||\mathcal{S}|)$ in the fully connected transition dynamics case, and $O(T(|\mathcal{A}| + |\mathcal{S}|))$ when the number of next states is bounded by a constant.

Note that Algorithm 9.1 and Algorithm 9.3 strictly generalize the forward-backward algorithm for efficient inference in Markov chain probability distributions (*e.g.*, hidden Markov models, conditional random fields).

Algorithm 9.3 Expected state frequency calculation

Require: MDP, $\mathcal{M}_{\text{MDP}}$, stochastic policy, $\pi(a_{x,y}|s_x)$), and initial state distribution $\text{P}_0(s_x)$.
Ensure: state visitation frequencies, D_{s_x} under policy $\pi(a_{x,y}|s_x)$.

for all states $s_x \in \mathcal{M}_{\text{MDP}}$ **do**
  $D_{s_x} \leftarrow 0$
end for
while not converged **do**
  for all states $s_x \in \mathcal{M}_{\text{MDP}}$ **do**
    $D'_{s_x} \leftarrow \text{P}_0(s_x)$
  end for
  for all actions $a_{x,y} \in \mathcal{M}_{\text{MDP}}$ **do**
    for all states $s_z \in \mathcal{M}_{\text{MDP}}$ reachable by $a_{x,y}$ **do**
      $D'_{s_z} \leftarrow D'_{s_z} + D_{s_x}\ \pi(a_{x,y}|s_x)\ \text{P}(s_z|a_{x,y}, s_x)$
    end for
  end for
  for all states $s_x \in \mathcal{M}_{\text{MDP}}$ **do**
    $D_{s_x} \leftarrow D'_{s_x}$
  end for
end while

9.1.2 Deterministic Dynamics Simplifications

In the setting with deterministic state transition dynamics (Section 6.2.5), the structure of the problem can be exploited with specific algorithms based on the reversibility of computing the partition function.

Theorem 9.1. *The expected action visitation frequencies, $D_{a_{x,y}}$, for origin s_a and goal state s_b in the deterministic dynamics maximum causal entropy statistic-matching model can be computed from partition functions (and log partition functions, $V_{s_a \to s_b}$, using the following equation:*

$$\begin{aligned} D_{a_{x,y}} &= \frac{Z_{s_a \to s_x}\ \ e^{\textit{reward}(a_{x,y})}\ \ Z_{s_y \to s_b}}{Z_{s_a \to s_b}} \\ &= e^{V_{s_a \to s_x} + \textit{reward}(a_{x,y}) + V_{s_y \to s_b} - V_{s_a \to s_b}} \end{aligned} \tag{9.4}$$

To calculate expected action frequencies, we therefore only need to compute the partition function (or log partition function) between all states and endpoints in our Markov decision process. We employ dynamic programming to recursively compute the log partition function forward from an initial state s_a in Algorithm 9.4. This algorithm has $O(T(|\mathcal{S}| + |\mathcal{A}|))$ run time, as a special case of the stochastic dynamics algorithm.

Algorithm 9.4 Forward log partition function calculation (deterministic)

Require: MDP, $\mathcal{M}_{\text{MDP}}$, initial and final endpoints s_a and s_b.
Ensure: $V^{\text{soft}}(s_a, s_x)$ is the log partition function from s_a to s_x from all $s_x \in \mathcal{S}$.

for all $s_x \in \mathcal{M}_{\text{MDP}}$ **do**
  $V^{\text{soft}}(s_a, s_x) \leftarrow -\infty$
end for
while not converged **do**
  for all $s_x \in \mathcal{M}_{\text{MDP}}$ **do**
    $V^{\text{soft}'}(s_a, s_x) \leftarrow 0$ if $s_a = s_x$ and $-\infty$ otherwise
  end for
  for all $a_{x \to y} \in \mathcal{M}_{\text{MDP}}$ **do**
    $V^{\text{soft}}(s_a, s_y)' \leftarrow \text{softmax}\left(V^{\text{soft}}(s_a, s_y)', \text{reward}_{\mathcal{M}_{\text{MDP}}}(s_x, a_{x \to y}) + V^{\text{soft}}(s_a, s_x)\right)$
  end for
  for all $s_x \in \mathcal{M}_{\text{MDP}}$ **do**
    $V^{\text{soft}}(s_a, s_x) \leftarrow V^{\text{soft}}(s_a, s_x)'$
  end for
end while

The resulting forward log partition function can be combined with the "backwards" log partition function obtained in Algorithm 9.1 to calculate the action visitation frequencies according to Equation 9.4.

Matrix-algebraic algorithm

A matrix-algebraic approach for exact inference in the deterministic dynamics setting makes use of the **geometric series of matrices** (Theorem 9.2).

Theorem 9.2. *For matrix* $\boldsymbol{A}$*, if* $\lim_{t \to \infty} \boldsymbol{A}^t = \boldsymbol{0}$*, then* $\sum_{t=0}^{\infty} \boldsymbol{A}^t = (\boldsymbol{I} - \boldsymbol{A})^{-1}$*, where* $\boldsymbol{I}$ *is the identity matrix and* $\boldsymbol{A}^0 = \boldsymbol{I}$*.*

Algorithm 9.5 makes use of this geometric series expression to compute the partition functions between all pairs of state endpoints.
The required matrix inversion (Step 5) can be performed by simple Gaussian elimination in $O(|\mathcal{S}|^3)$ time or by the Coppersmith-Winograd algorithm (Coppersmith & Winograd, 1990) in $O(|\mathcal{S}|^{2.376})$ time[1]. However, if matrix **B** is very large, but also sparse (*i.e.*, the total number of actions is much smaller than $|\mathcal{S}|^2$), power-iteration methods are known to have better practical performance (*e.g.*, the PageRank algorithm (Brin & Page, 1998)).

[1]The $O(|\mathcal{S}|^{2.376})$ is largely a theoretical guarantee; algorithms with worse asymptotic run time are more efficient in practice on matrices of practical consideration.

Algorithm 9.5 Partition function calculation via matrix inversion

Require: MDP, $\mathcal{M}_{\text{MDP}}$
Ensure: $\{Z_{A \to B} = \sum_{\zeta_{A \to B}} e^{\text{reward}_{\mathcal{M}_{\text{MDP}}}(\zeta_{A \to B})}\}$.
1: $\mathbf{A} \leftarrow \mathbf{0}$
2: **for all** actions $a_{X \to Y} \in \mathcal{M}_{\text{MDP}}$ **do**
3: $\quad \mathbf{A}_{X,Y} \leftarrow e^{\text{reward}_{\mathcal{M}_{\text{MDP}}}(S_X) + \text{reward}_{\mathcal{M}_{\text{MDP}}}(A_{X \to Y})}$
4: **end for**
5: $\mathbf{B} \leftarrow (\mathbf{I} - \mathbf{A})^{-1} - \mathbf{I}$
6: **for all** state pairs $S_A \in \mathcal{M}_{\text{MDP}}$ and $S_B \in \mathcal{M}_{\text{MDP}}$ **do**
7: $\quad Z_{A \to B} \leftarrow \mathbf{B}_{A,B}$
8: **end for**

9.1.3 Convergence Properties and Approximation

All of the algorithms that we have presented for inference require convergence—either of the values of a dynamic program (Algorithm 9.1), the steady-state visitation values (Algorithm 9.3), or the sequence of matrix multiplications (Algorithm 9.5). For finite time horizon settings, convergence is guaranteed. However, assuming a finite horizon may be inappropriate for many application settings. The natural question is then: in which settings do these algorithm convergence requirements hold over infinite horizons?

Remark 9.3. *The convergence properties of the maximum causal entropy model can be broken into three regimes:*

- ***Strong non-convergence:*** *the most likely policy under the model produces state-action sequences that, in expectation, have infinite length.*
- ***Weak non-convergence:*** *the most likely policy produces state-action sequences that are, in expectation, finite in length, but the expected length of all state-action sequences (averaged over all policies) is infinite in length.*
- ***Convergence:*** *expected state-action sequence lengths are finite when generated under the most likely policy and when averaged over all policies.*

Strong non-convergence has a straight-forward parallel in deterministic planning problems: when a cycle of positive reward exists along a path between two endpoints, the optimal plan (and the log partition function) will be undefined and its value will not converge. For stochastic state-transition dynamics, the analogy is having a subset of states where a policy that only transitions to that subset of states yields positive expected reward from any state from that subset. Weak non-convergence is specific to the probabilistic, softened Bellman setting, and has no optimal control parallel. Due to this difference, restraints that guarantee optimal control convergence do not guarantee convergence in the maximum causal entropy model.

Remark 9.4. *Though strictly negative rewards ensure convergence in the state and action value functions of the optimal control setting, they do not ensure convergence in the corresponding maximum causal entropy setting (i.e., of the soft-maximum log partition functions for states and actions).*

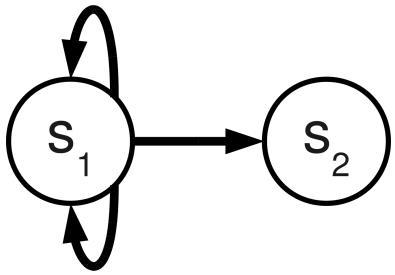

Figure 9.1: An illustrative example of non-convergence with strictly negative rewards for each deterministic action. We consider the case where each action has a cost (*i.e.*, a negative reward) $\epsilon > 0$. Note that the most likely state-action sequence is the immediate transition to from s_1 to s_2 (with cost ϵ). However, the contribution to the partition function for paths that cycle at state s_1 a total of x times is $2^x e^{-x\,\epsilon} = e^{x \ln 2 - x\,\epsilon}$. Thus, for $\epsilon \leq \ln 2$, the partition function does not converge.

Figure 9.1 provides a simple example demonstrating the non-convergence of the potential function in the weak non-convergence regime of behavior (Remark 9.4).

A number of structural restrictions on the state transition dynamics do ensure partition function convergence in general decision settings. We note three:

- **Directed, acyclic decision structures:** the state transition dynamics restrict the sequence of states and actions so that no sequence can revisit the same state. Thus, every trajectory must terminate with finite length, guaranteeing partition function convergence.

- **Finite horizon structures:** finite length trajectory is considered with potentially cyclic sequences. Explicit dependence on the time step is required by the log partition function:

$$V_{s_x \to s_b,\, t+1} = \operatorname*{softmax}_{a_{x,y}} \left(\sum_{s_z} \mathrm{P}(s_z | a_{x,y}, s_x) V_{s_y \to s_b,\, t} \right), \tag{9.5}$$

  and in the corresponding time-varying policy.

- **Discounted future rewards:** at each timestep the state-action sequence ends with probability $(1 - \gamma)$ for $\gamma > 0$. The remaining outcomes dynamics are multiplicatively scaled by the discount factor, γ.

An important convergence guarantee results from employing discounting.

Theorem 9.5. *In a discounted future rewards setting with a bounded instantaneous reward and a bounded number of actions, the partition functions of the maximum causal entropy model with finite rewards are guaranteed to converge.*

Convergence is also possible in infinite horizon, non-discounted, cyclic decision structures, but only for some choices of parameters. For example, in the deterministic setting when the eigenvalues of the constructed transition reward matrix (Algorithm 9.5) are less than 1.

9.1.4 Propagation Optimizations and Approximations

For large MDPs, Algorithms 9.4 and 9.1 may be too slow in practice to employ for computing the partition functions for the class of all paths of a large length (T). We could just consider a smaller length, but small length paths may represent only a small portion of the probability mass in the complete set of paths or fail to connect from constrained origin and terminal states. For example, under certain parameter choices, a parameterized reward Markov decision process assigns nearly all probability mass to the optimal action choices of each state. Intuitively, approaches similar to the specialized, fast algorithms for finding optimal solutions for decision problems (*e.g.*, A* search (Hart et al., 1968) for planning settings) should be employed for the "softened" version of those decision settings as well. Based on this motivation, we consider a smaller class of paths that contains most of the probability mass of the larger class of paths, while being significantly more efficient to compute.

Definition 9.6. *Consider some ordering, Order$(S) \rightarrow \mathbb{Z}$, over states of the Markov decision process. We call a path, ζ, **consistent** with this ordering if $\forall_{S_t,S_{t+1}\in\zeta}$ Order$(S_t) <$ Order(S_{t+1}). We call a path **k-inconsistent** with ordering Order if $\forall_{S_t,S_{t+1}\in\zeta}$ Order$(S_t) >$ Order(S_{t+1}) at most k time.*

Rather than consider all paths of length T in our inference, we will now consider all paths that are k-consistent with an ordering.

If we have convergent partition functions, and consider a very large length T and consistency k, the set of paths that differ between the two classes will be small and contribute only a very small amount to the overall probability mass of log partition functions.

To help illustrate the advantages of this class of paths, consider states s_a and s_b in a Markov decision process with the minimum length of paths connecting the two states being L. We will have to consider paths of length greater than L using more than L iterations of our standard inference algorithm to have any estimate at all for visitation frequencies. However, if the ordering is carefully chosen, the order-consistent set of paths will contain some paths from s_a to s_b and provide a reasonable estimate of visitation frequencies.

We can efficiently consider the log partition functions based on the k-consistent class of paths for some ordering by employing the dynamic program of Algorithm 9.6.

State log partition values are updated according to the ordering and, unlike the non-optimized variant (Algorithm 9.1), the results of the updates from earlier in the ordering are propagated

Algorithm 9.6 Optimized stochastic policy calculation

Require: MDP, $\mathcal{M}_{\text{MDP}}$, state ordering, Order, and terminal state reward function, $\phi(s) \rightarrow \mathbb{R}$.
Ensure: a maximum causal entropy policy, $\pi(a_{x,y}|s_x)$.

for all States s_x **do**
  $V_{s_x} \leftarrow -\infty$
end for
$t \leftarrow 0$
while not converged **do**
  for all States s_x according to the reverse of ordering Order **do**
    $V_{s_x} \leftarrow \phi(s_x)$
    for all Actions $a_{x,y}$ from state s_x **do**
      $V_{s_x} \leftarrow \text{softmax}\left(V_{s_x}, \sum_Z \text{P}(s_z|s_x, a_{x,y})V_{s_z}\right)$
    end for
  end for
  for all Actions $a_{x,y}$ **do**
    $\pi_t(a_{x,y}|s_x) \leftarrow e^{\left(\sum_z \text{P}(s_z|s_x,a_{x,y})V_{s_z}\right)-V_{s_x}}$
  end for
  $t \leftarrow t + 1$
end while

through states later in the ordering. In the deterministic dynamics setting, with state action sequences constrained to start in state s_a and terminate in state s_b, we employ a heuristic to order each state (s_i) based on the state's fractional **progress** between s_a and s_b. Three heuristics based on this notion of progress are:

$$\text{progress}_1(s_i) = \frac{\min_{\zeta_{s_a \rightarrow s_i}} \text{reward}(\zeta_{s_a \rightarrow s_i})}{\min_{\zeta_{s_a \rightarrow s_i}} \text{reward}(\zeta_{s_a \rightarrow s_i}) + \min_{\zeta_{s_i \rightarrow s_b}} \text{reward}(\zeta_{s_i \rightarrow s_b})}; \tag{9.6}$$

$$\text{progress}_2(s_i) = 1 - \frac{\min_{\zeta_{s_i \rightarrow s_b}} \text{reward}(\zeta_{s_a \rightarrow s_i})}{\min_{\zeta_{s_a \rightarrow s_i}} \text{reward}(\zeta_{s_a \rightarrow s_i}) + \min_{\zeta_{s_i \rightarrow s_b}} \text{reward}(\zeta_{s_i \rightarrow s_b})}; \tag{9.7}$$

$$\text{progress}_3(s_i) = \left\{ \begin{array}{ll} \text{progress}_1(s_i) & \text{if } \min_{\zeta_{s_a \rightarrow s_i}} < \min_{\zeta_{s_i \rightarrow s_b}} \\ \text{progress}_2(s_i) & \text{otherwise} \end{array} \right\} \tag{9.8}$$

We then generate our state order based on this measure of progress using a fixed domain-dependent reward function. When there is no specific terminal state constraint, the cost of the optimal path from the initial state, $\min_{\zeta_{s_a \rightarrow s_i}} \text{reward}(\zeta_{s_a \rightarrow s_i})$, can be employed as an ordering criteria for the states. Similar heuristic-based approaches can be employed in the stochastic setting by first solving a fixed parameter optimal control problem and employing the resulting state value function to order the states.

The expected state and action visitation frequencies are similarly calculated according to this optimized reward function using Algorithm 9.7.

Algorithm 9.7 Optimized expected state frequency calculation

Require: MDP, $\mathcal{M}_{\text{MDP}}$, time-varying policy, $\pi_t(a_{x,y}|s_x))$, and an initial state distribution $\mathrm{P}_0(s_x)$.
Ensure: state visitation frequencies, D_{s_x}, under policy $\pi_t(a_{x,y}|s_x)$.

for all states $s_x \in \mathcal{M}_{\text{MDP}}$ **do**
  $D_{s_x} \leftarrow 0$
end for
$t \leftarrow$ largest time-index of π
while not converged **do**
  for all states s_z according to ordering O **do**
    $D_{s_z} \leftarrow \mathrm{P}_0(s_z)$
    for all actions $a_{x,y} \in \mathcal{M}_{\text{MDP}}$ such that S_z is a possible stochastic outcome of $a_{x,y}$ **do**
      $D_{s_z} \leftarrow D_{s_z} + D_{s_x}\ \pi_t(a_{x,y}|s_x)\ \mathrm{P}(s_z|a_{x,y}, s_x)$
    end for
  end for
  $t \leftarrow t - 1$
end while

As a second optimization, paths through states that are very far away from A and B often contribute little to the probability mass of the partition function. We only include states in our ordering that are within a certain threshold: $\min_{\zeta_{A\to i}} cost(\zeta_{A\to i}) + \min_{\zeta_{i\to B}} cost(\zeta_{i\to B}) < C$, where the threshold is determined by $\min_{\zeta_{A\to B}} cost(\zeta_{A\to B})$. For example, $C = \alpha \min_{\zeta_{s_a\to s_i}} \text{cost}(\zeta_{s_a\to s_i}) + \epsilon$ for some $\alpha > 1$ and $\epsilon > 0$. Optimal state values (*i.e.*, optimal path costs in the deterministic setting) are again based on a fixed heuristic.

While additional optimizations for the inference algorithm are possible based on intelligently ordering the updates, there is an important technical consideration to keep in mind: any heuristics employed to guide inference optimizations that depends on a learned cost weight may create non-convexity when trying to optimize for that cost weight. In other words, adapting the ordering can lead to non-convexity when fitting the parameters of a maximum causal entropy model to observed data. This risk may be appropriately mitigated—either in theory or in practice—but we do not explore dynamic re-orderings in this work.

9.1.5 Linear Quadratic Inference

For continuous state and continuous control inverse optimal control, the inference procedure is very similar to Algorithm 9.1, but instead relies on the specific properties of Gaussian distributions and quadratic reward functions to obtain closed-form expressions. As established by Theorem 6.15, when the state transition dynamics follow a linear function, $\mathbf{s}_{t+1} \sim N(\mathbf{A}\mathbf{s}_t + \mathbf{B}\mathbf{a}_t, \Sigma)$ for given matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{\Sigma}$, the time-varying state-action and state values are:

$$Q_\theta^{\text{soft}}(\mathbf{a}_t, \mathbf{s}_t) = \begin{bmatrix} \mathbf{a}_t \\ \mathbf{s}_t \end{bmatrix}^\top \begin{bmatrix} \mathbf{B}^\top\mathbf{D}\mathbf{B} & \mathbf{A}^\top\mathbf{D}\mathbf{B} \\ \mathbf{B}^\top\mathbf{D}\mathbf{A} & \mathbf{A}^\top\mathbf{D}\mathbf{A} \end{bmatrix} \begin{bmatrix} \mathbf{a}_t \\ \mathbf{s}_t \end{bmatrix} + \begin{bmatrix} \mathbf{a}_t \\ \mathbf{s}_t \end{bmatrix}^\top \begin{bmatrix} \mathbf{B}^\top\mathbf{G} \\ \mathbf{A}^\top\mathbf{G} \end{bmatrix}$$
$$V_\theta^{\text{soft}}(\mathbf{s}_t) = \mathbf{s}_t^\top(\mathbf{C}_{s,s} + \mathbf{Q} - \mathbf{C}_{a,s}^\top\mathbf{C}_{a,a}^{-1}\mathbf{C}_{a,s})\mathbf{s}_t + \mathbf{s}_t^\top(\mathbf{F}_s + \mathbf{R}) + \text{const},$$

where $\mathbf{C}$ and $\mathbf{D}$ are recursively computed as: $\mathbf{C}_{a,a} = \mathbf{B}^\top\mathbf{D}\mathbf{B}; \mathbf{C}_{s,a} = \mathbf{C}_{a,s}^\top = \mathbf{B}^\top\mathbf{D}\mathbf{A}; \mathbf{C}_{s,s} = \mathbf{A}^\top\mathbf{D}\mathbf{A}; \mathbf{D} = \mathbf{C}_{s,s} + \mathbf{Q} - \mathbf{C}^\top\mathbf{C}_{a,a}^{-1}\mathbf{C}_{a,s}$; and $\mathbf{G} = \mathbf{F}_s + \mathbf{R}$.

Given terminal state quadratic reward weights, $\phi(\mathbf{s}) = \mathbf{s}^\top\mathbf{Q}_{\text{terminal}}\mathbf{s} + \mathbf{s}^\top\mathbf{R}$ and instantaneous reward weights, $\mathbf{Q}_{\text{instant}}$ and $\mathbf{R}_{\text{instant}}$, Algorithm 9.8 generates time-varying quadratic state-value and action-value parameters.

Algorithm 9.8 Linear-quadratic regulation value inference.

Require: Linear-quadratic model, $\mathcal{M}_{\text{lqr}}$, with terminal quadratic reward parameters, $\mathbf{Q}_{\text{terminal}}$ and $\mathbf{R}_{\text{terminal}}$; and time horizon, T.

Ensure: Time-varying action-value and state-value quadratic parameters: $\mathbf{C}$ and $\mathbf{D}$; and linear parameters: $\mathbf{F}$ and $\mathbf{G}$.

$\mathbf{D}_0 \leftarrow \mathbf{Q}_{\text{terminal}}$
$\mathbf{G}_0 \leftarrow \mathbf{R}_{\text{terminal}}$
for $t = 1$ to T **do**
  $\mathbf{C}_{a,a} \leftarrow \mathbf{B}^\top\mathbf{D}_{t-1}\mathbf{B}$
  $\mathbf{C}_{s,a} \leftarrow \mathbf{B}^\top\mathbf{D}_{t-1}\mathbf{A}$
  $\mathbf{C}_{a,s} \leftarrow \mathbf{C}_{s,a}{}^\top$
  $\mathbf{C}_{s,s} \leftarrow \mathbf{A}^\top\mathbf{D}_{t-1}\mathbf{A}$
  $\mathbf{C}_t \leftarrow \begin{bmatrix} \mathbf{C}_{a,a} & \mathbf{C}_{s,a} \\ \mathbf{C}_{a,s} & \mathbf{C}_{s,s} \end{bmatrix}$
  $\mathbf{G}_t \leftarrow \mathbf{F}_s + \mathbf{R}_{\text{instant}}$
  $\mathbf{D}_t \leftarrow \mathbf{C}_{s,s} + \mathbf{Q}_{\text{instant}} - \mathbf{C}_{s,a}\mathbf{C}_{a,a}^{-1}\mathbf{C}_{a,s}$
  $\mathbf{F}_a \leftarrow \mathbf{B}^\top\mathbf{G}$
  $\mathbf{F}_s \leftarrow \mathbf{A}^\top\mathbf{G}$
  $\mathbf{F}_t \leftarrow \begin{bmatrix} \mathbf{F}_a \\ \mathbf{F}_s \end{bmatrix}$
end for

The state and action matrix multiplications dominate the run time of Algorithm 9.8. With naïve matrix multiplication, the run time is $O(T(\dim(\mathbf{S})^3 + \dim(\mathbf{A})^3))$.

The stochastic policy is then obtained from the state-action parameter, $\mathbf{C}$:

$$\pi(\mathbf{a}|\mathbf{s}) \propto e^{-[\mathbf{a}\ \mathbf{s}]^\top\mathbf{C}[\mathbf{a}\ \mathbf{s}] - [\mathbf{a}\ \mathbf{s}]^\top\mathbf{F}}, \tag{9.9}$$

which is a multi-variate Gaussian distribution.

Algorithm 9.9 Linear-quadratic state and action distribution calculation.

Require: Linear-quadratic model, $\mathcal{M}_{\text{lqr}}$, stochastic policy parameter, $\mathbf{C}$, and initial state, $\mathbf{s}_0$.
Ensure: A sequence of state and action distributions, $\mathrm{P}(\mathbf{s}_t)$ and $\mathrm{P}(\mathbf{a}_t)$.
1: **for** $t = 1$ to T **do**
2: Compute joint distribution $\mathrm{P}(\mathbf{a}_t, \mathbf{s}_t)$ from $\mathrm{P}(\mathbf{s}_t)$ and $\pi(\mathbf{a}_t|\mathbf{s}_t)$
3: Compute $\mathrm{P}(\mathbf{a}_t)$ by marginalizing over $\mathbf{s}_t$: $\mathrm{P}(\mathbf{a}_t) = \int_{\mathbf{s}_t} \mathrm{P}(\mathbf{a}_t, \mathbf{s}_t)\ \partial \mathbf{s}_t$
4: Compute joint distribution $\mathrm{P}(\mathbf{s}_{t+1}, \mathbf{a}_t, \mathbf{s}_t)$ from $\mathrm{P}(\mathbf{a}_t, \mathbf{s}_t)$ and $\mathrm{P}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$
5: Compute $\mathrm{P}(\mathbf{s}_{t+1})$ by marginalizing over $\mathbf{a}_t$ and $\mathbf{s}_t$: $\mathrm{P}(\mathbf{s}_{t+1}) = \int_{\mathbf{s}_t, \mathbf{a}_t} \mathrm{P}(\mathbf{s}_{t+1}, \mathbf{a}_t, \mathbf{s}_t)\ \partial \mathbf{s}_t\ \partial \mathbf{a}_t$
6: **end for**

We compute the joint distribution of states and actions from the state-dependent stochastic policy by employing Algorithm 9.9.

Remark 9.7. *The marginalized state and actions distributions are obtained from the following linear relationships:*

$$P(\boldsymbol{s}_t) \sim N(\boldsymbol{s}_t|\mu_{s_t}, \Sigma_{s_t}) \tag{9.10}$$

$$P(\boldsymbol{a}_t|\boldsymbol{s}_t) \sim N(\boldsymbol{a}_t| - C_{a,a}^{-1} C_{a,s} s_t, \sigma_{a,a}) \tag{9.11}$$

$$P(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t) \sim N(\boldsymbol{s}_{t+1}|A\ \boldsymbol{s}_t + B\ \boldsymbol{a}_t, \Sigma_{dyn}), \tag{9.12}$$

by iteratively conditioning the Gaussian distributions and then marginalizing over previous variables.

Joint distributions and marginal distributions are obtained by transforming between the standard form (i.e., mean μ and variance Σ) and the canonical, quadratic form:

$$\mathrm{P}(\boldsymbol{x}) \propto e^{\eta^\top \boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^\top \Lambda \boldsymbol{x}}, \tag{9.13}$$

which are related by: $\mu = \Lambda^{-1}\eta$ and $\Sigma = \Lambda^{-1}$. Marginalization is simpler in standard form (by taking a sub-matrix of Σ and sub-vector of μ), while conditioning is simpler to apply in canonical, quadratic form (by addition of the precision matrix, Λ and the vector η).

9.2 Latent Information Inference

We now consider the maximum causal entropy inference problems for influence diagrams. This decision framework generalizes the inverse optimal control setting by allowing latent side information variables and imperfect side information recall. As the maximum causal entropy approach is only valid when duality holds, *i.e.*, when perfect decision recall is guaranteed, we consider inference within this setting.

9.2.1 Perfect Recall Visitation Counts

We start by considering a stronger restriction on the influence diagram. For the special case of perfect recall of both past decisions and past conditioning variables (Definition 7.9), reducing the latent information inference to the inverse optimal control inference procedure is possible by marginalizing and employing expectations.

Algorithm 9.10 MaxCausalEnt ID inference procedure for perfect recall

Require: A maximum causal entropy influence diagram with perfect recall.
Ensure: Partition functions $Z(Y_i|\text{par}(Y_i))$ that are maximum causal entropy distributed.

1: **for all** $V \in \mathbf{V}$ **do**
2: Associate V with $\text{argmax}_{Y_{\text{index}} \in \text{ancest}(V)}$ index
3: **end for**
4: **for** $i = |\mathbf{Y}|$ to 1 **do**
5: **for all** values $(Y_i', \text{par}(Y_i)')$ **do**
6: $Z_i(Y_i'|\text{par}(Y_i)') \leftarrow 0$
7: **for all** V_j associated with Y_i **do**
8: **for all** values $\text{par}(V_j)'$ **do**
9: Compute $\text{P}(\text{par}(V_j)'|\text{par}(Y_i)', Y_i')$
10: **end for**
11: $Z_i(Y_i'|\text{par}(Y_i)') \leftarrow Z_i(Y_i'|\text{par}(Y_i)') + \text{E}_{\text{par}(V_j)}[\theta^\top \mathcal{F}_{V_j}(\text{par}(V_j))|Y_i', \text{par}(Y_i)']$
12: **end for**
13: **for all** values $\text{par}(Y_{i+1})'$ **do**
14: Compute $\text{P}(\text{par}(Y_{i+1})'|Y_i', \text{par}(Y_i)')$
15: **end for**
16: $Z_i(Y_i'|\text{par}(Y_i)') \leftarrow Z_i(Y_i'|\text{par}(Y_i)') + \text{E}_{\text{par}(Y_{i+1})'}[Z_{i+1}(\text{par}(Y_{i+1})')|Y_i', \text{par}(Y_i)']$
17: **end for**
18: **for all** $\text{par}(Y_i)'$ **do**
19: $Z_i(\text{par}(Y_i)') \leftarrow \text{softmax}_{Y_i} Z_i(Y_i|\text{par}(Y_i)')$
20: **end for**
21: **end for**

Algorithm 9.10 illustrates the procedure for inferring decision probabilities in this special case based on Theorem 6.2. Essentially, the dynamics of side information are calculated by marginalization and the expectations of values are computed. We assume as a subroutine an algorithm for calculating the marginal probabilities of variables conditioned on a set of fixed evidence variables in a Bayesian network (*e.g.*, variable elimination or belief propagation). Expectations over the unobserved uncertainty nodes that are ancestors of value variables are employed in line 11. This replaces the exact evaluations, $\theta^\top \mathcal{F}(\mathbf{X}, \mathbf{Y})$, of Equation 6.1.

9.2.2 Imperfect Recall Visitation Counts

In the imperfect recall inference problem with perfect recall of past actions, action distributions are recursively inter-dependent. As a result, the backward inductive dynamic programming technique of Algorithm 9.10 can not be employed since $\mathrm{P}(a_T|s_T)$ is not independent from the distribution of past actions given state s_T. Instead, we rely on a fixed point iteration algorithm that iteratively updates the stochastic policy in hope of obtaining a fixed point solution to the inter-dependent probability distributions.

Algorithm 9.11 MaxCausalEnt ID inference procedure for imperfect side information recall

Require: A maximum causal entropy influence diagram with perfect decision recall.
Ensure: Partition functions $Z(Y_i|\text{par}(Y_i))$ that are maximum causal entropy distributed.

for all actions nodes, a_t **do**
  Set $\mathrm{P}(a_t|\text{par}(a_t))$ to a uniform distribution
end for
$Z_{\text{par}(a_{T+1})} = \text{Utility}(\mathbf{s}, \mathbf{a})$
while not sufficiently converged **do**
  for $\tau = T$ to 1 **do**
    Compute $\mathrm{P}(\text{anc}(a_\tau)|\text{par}(a_\tau))$
    Compute $\mathrm{P}(\text{par}(a_{\tau+1})|a_\tau, \text{par}(a_\tau))$
    for all $a_\tau, \text{par}(a_\tau)$ **do**
      $Z_{a_\tau|\text{par}(a_\tau)} = \mathrm{E}_{\mathrm{P}(\text{par}(a_{\tau+1})|a_\tau,\text{par}(a_\tau))}[Z_{\text{par}(a_{\tau+1})}|a_\tau, \text{par}(a_\tau)]$
    end for
    for all $\text{par}(a_\tau)$ **do**
      $Z_{\text{par}(a_\tau)} = \mathrm{softmax}_{a_\tau} Z_{a_\tau|\text{par}(a_\tau)}$
    end for
  end for
end while

Algorithm 9.11 illustrates the procedure for obtaining the maximum causal entropy policy for this more general imperfect information setting with perfect past decision recall.

9.3 Regret-Based Model Inference

The algorithms for considering multi-agent game settings maximize the causal entropy of the joint distribution of players' actions at each point in time, denoted $a^t_{1:N}$, while enforcing rationality constraints.

9.3.1 Correlated Equilibria Inference

As established by Theorem 8.15, maximum causal entropy correlated equilibria are distributed according to Equation 9.14:

$$\pi_\lambda^{\text{MCECE}}(a_{1:N}^t|s^t) \propto e^{-\left(\sum_{i,a_i^{t\prime}} \lambda_{i,s^t,a_i^t,a_i^{t\prime}} \text{ExpectDevGain}_i^{\pi_\lambda^{\text{MCECE}}}\left(a_{1:N}^t,s^t,a_i^{t\prime}\right)\right)+\text{ExpectEnt}_\lambda\left(a_{1:N}^t,s^t\right)}, \tag{9.14}$$

where $\text{ExpectEnt}(a_{1:N}^t, s^t) \triangleq \mathrm{E}_{a_{1:N}^{t+1},s^{t+1}}\left[\text{ExpectEnt}_\lambda(a_{1:N}^{t+1}, s^{t+1}) + H(a_{1:N}^{t+1}|s^{t+1})\middle|a_{1:N}^t, s^t\right]$. The free parameters, λ, are chosen to maximize the causal entropy while enforcing deviation regret constraints for the joint action distribution of each state.

The naïve approach to this problem is to jointly maximum the entropy of the optimization to find all λ parameters for each time step and state in parallel, as shown in Algorithm 9.12.

Algorithm 9.12 MCECE strategy profile computation for finite horizon

Require: A fully-specified Markov game.
Ensure: A maximum causal entropy correlated equilibrium.
1: $\lambda^{(1)} = \{\lambda^{(1)}_{t,i,s^t,a_i^t,a_i^{t\prime}}\} \leftarrow$ (arbitrary) positive initial values.
2: $x \leftarrow 1$
3: **while** not converged **do**
4: Compute $\pi_\lambda^{(x)} = \{\pi_\lambda(a_{1:N}^t|s^t)\}$ from $\lambda^{(x)}$ using a subroutine.
5: Take an optimization step using $\pi_\lambda^{(x)}$ to improve $\lambda^{(x+1)}$
6: $x \leftarrow x + 1$
7: **end while**

However, the Markovian relationship between variables can be exploited to improve the optimization procedure.

Remark 9.8. *For time-varying policies, future strategy probabilities (and dual parameters) are independent of earlier strategy and dual parameters given the state. As a result, a sequence dynamic programming approach can be employed to compute the maximum causal entropy equilibrium.*

Following Remark 9.8, Algorithm 9.12 can be re-expressed as a sequential dynamic programming algorithm (Algorithm 9.13) resembling value iteration (Bellman, 1957). It iteratively computes both future expected utilities and expected entropies by fully optimizing λ parameters at timestep T and then employing the corresponding expected utilities and expected entropies for the optimization at timestep $T-1$. This process is iteratively continued to obtain the policy for all timesteps.

We investigate algorithms for the optimization problem of Step 4 of Algorithm 9.13 in Chapter 10.

Algorithm 9.13 Value iteration approach for obtaining MCECE

Require: A fully-specified Markov game.
Ensure: A maximum causal entropy correlated equilibrium.

$\forall_{i,a_{1:N},s}$ $\text{ExpectUtil}_i(a_{1:N}, s) \leftarrow \text{Utility}_i(a_{1:N}, s)$
$\forall_{a_{1:N},s}$ $\text{ExpectEnt}(a_{1:N}, s) \leftarrow 0$
for $t = T$ to 1 **do**
  For each state, s^t, obtain $\{\pi_\lambda(a^t_{1:N}|s^t)\}$ using ExpectUtil and ExpectEnt values in the following optimization:

$$\underset{\pi(a^t_{1:N}|s^t)}{\operatorname{argmax}} H(a^t_{1:N}|s^t) + \mathrm{E}_{\pi(a^t_{1:N}|s^t)}\left[\text{ExpectEnt}(a_{1:N}, s)|s^t\right]$$

$$\text{such that: } \sum_{a_{-i}} \mathrm{P}(a^t_{1:N}|s^t)\left(\text{ExpectRegret}(\{a^t_{-i}, a^t_i\}, s^t) - \text{ExpectRegret}(a^t_{1:N}, s^t)\right) \leq 0$$

$$\forall_{a^t_{1:N}} \mathrm{P}(a^t_{1:N}|s^t) \geq 0 \text{ and } \sum_{a^t_{1:N}} \mathrm{P}(a^t_{1:N}|s^t) = 1.$$

  $\forall_{i,a_{1:N},s}$ $\text{ExpectUtil}'_i(a_{1:N}, s) \leftarrow \gamma \sum_{a^t_{1:N}, s^t} \pi(a^t_{1:N}|s^t)\,\mathrm{P}(s^t|s, a_{1:N})\,\text{ExpectUtil}_i(a^t_{1:N}, s^t)$
  $\forall_{s,a_{1:N}}$ $\text{ExpectEnt}'(a_{1:N}, s) \leftarrow \gamma \sum_{a^t_{1:N}, s^t} \pi(a^t_{1:N}|s^t)\,\mathrm{P}(s^t|s, a_{1:N})\big(\text{ExpectEnt}(a^t_{1:N}, s^t) + H(a^t_{1:N}|s^t)\big)$
  $\forall_{i,a_{1:N},s}$ $\text{ExpectUtil}_i(a_{1:N}, s) \leftarrow \text{ExpectUtil}'_i(a_{1:N}, s) + \text{Utility}_i(a_{1:N}, s)$
  $\forall_{a_{1:N},s}$ $\text{ExpectEnt}(a_{1:N}, s) \leftarrow \text{ExpectEnt}'(a_{1:N}, s)$
end for

9.4 Discussion

In this chapter, we have presented algorithms for efficient inference within the maximum causal entropy framework. For equality-constrained maximum causal entropy, the inference algorithms are softened extensions of value iteration dynamic programming procedures. With inequality-based rationality constraints, a similar dynamic programming procedure is employed to propagate information backwards over time. However, a convex optimization is required to obtain the maximum causal entropy policy at each state and timestep.

The efficiency provided by these algorithms is important. Rather than reasoning depending on an entire history, which grows exponentially with history length, reasoning is generally linear in the history size, as long as the dynamics governing side information are Markovian. This efficiency enables the maximum causal entropy approach to be applied to large decision prediction tasks in Part IV of the thesis.

Chapter 10

Parameter Learning

"Our knowledge is a little island in a great ocean of non-knowledge."
— Isaac Singer (Inventor, 1811–1875).

Accurately predicting future behavior based on previously observed behavior sequences requires fitting the parameters of the maximum causal entropy probability distribution using that observed training data. Since no closed-form solution exists for the best parameter choice in general, maximum entropy optimization-based approaches must instead be employed to obtain parameters that provide accurate prediction. In this chapter, we describe the maximum causal entropy gradients and the gradient-based optimization methods that we employ to learn model parameters from training data. In the case of rationality requirements for behavior in sequential games, optimizations techniques are required to enforce the rationality constraints independently of observed game play.

10.1 Maximum Causal Entropy Model Gradients

By employing the Lagrangian of the maximum causal entropy optimization (Definition 5.6) and solving for the causally conditioned probability distribution form, we are left to only enforce the remaining constraints—either purposeful equality constraints or inequality constraints guaranteeing rationality.

10.1.1 Statistic-Matching Gradients

In the statistic-matching maximum causal entropy model (Chapter 6), the gradient of the dual is the difference between empirical feature counts and expected feature counts:

$$\nabla_\theta \, F_{\text{dual}}(\theta) = \left(\mathrm{E}_{\tilde{\mathrm{P}}(\mathbf{X},\mathbf{Y})} \left[\sum_t F(x_t, y_t) \right] - \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})} \left[\sum_t F(x_t, y_t) \right] \right). \tag{10.1}$$

We employ Algorithm 10.1 to compute the expected feature sum under the maximum causal entropy model's policy (obtained via Algorithm 9.1). The empirical feature counts can be trivially obtained by calculating the empirical visitation frequency and weighting each state's feature by this count.

Algorithm 10.1 Feature expectation calculation

Require: MDP $\mathcal{M}_{\text{MDP}}$, initial state s_0, stochastic policy, $\pi(a|s)$.
Ensure: Expected feature counts, $\mathrm{E}[\mathbf{f}]$, under the stochastic policy, $\pi(a|s)$.

```
Compute D_{s_x} under π(a|s) from state s_0 using Algorithm 9.7
E[f] ← 0
for all s_x ∈ M_MDP do
    for all a_i from state s_x do
        E[f] ← E[f] + D_{s_x} π(a_i|s_x) f_{s_x,a_i}
    end for
end for
```

In the linear-quadratic regulation setting, expected sums of quadratic states and actions serve as continuous analogs to the feature function sums of the discrete state and action setting. The maximum causal entropy dual's gradient in this setting is simply the difference between these sums of quadratics or linear state functions under the empirical distribution and under the model's expectation, as shown in Equation 10.2 and Equation 10.3:

$$\frac{\partial F_{\text{dual}}(\theta)}{\partial \mathbf{Q}} = \left(\mathrm{E}_{\tilde{\mathrm{P}}(\mathbf{s}_{1:T},\mathbf{a}_{1:T})} \left[\sum_t \mathbf{s}_t \, \mathbf{s}_t^\top \right] - \mathrm{E}_{\mathrm{P}(\mathbf{s}_{1:T},\mathbf{a}_{1:T})} \left[\sum_t \mathbf{s}_t \, \mathbf{s}_t^\top \right] \right); \tag{10.2}$$

$$\frac{\partial F_{\text{dual}}(\theta)}{\partial \mathbf{R}} = \left(\mathrm{E}_{\tilde{\mathrm{P}}(\mathbf{s}_{1:T},\mathbf{a}_{1:T})} \left[\sum_t \mathbf{s}_t \right] - \mathrm{E}_{\mathrm{P}(\mathbf{s}_{1:T},\mathbf{a}_{1:T})} \left[\sum_t \mathbf{s}_t \right] \right). \tag{10.3}$$

Employing the inferred state and action distributions at each point in time (Algorithm 9.9), the quadratic expectations of states and actions are calculated according to Algorithm 10.2.

The integrals of Algorithm 10.2 are computed using the following Gaussian property: $\mathrm{E}_{\mathbf{x} \sim N(\mu,\Sigma)}[\mathbf{x} \, \mathbf{x}^\top] = \mu\mu^\top + \Sigma$.

Algorithm 10.2 Quadratic expectation calculation

Require: LQR $\mathcal{M}_{\text{LQR}}$, state distribution $\text{P}(\mathbf{s}_t)$, action distribution $\text{P}(\mathbf{a}_t)$.
Ensure: Expectations, $\text{E}[\sum_t \mathbf{s}_t\ \mathbf{s}_t^\top]$ and $\text{E}[\sum_t \mathbf{a}_t\ \mathbf{a}_t^\top]$ based on distributions $\text{P}(\mathbf{s}_t)$ and $\text{P}(\mathbf{a}_t)$.
$\text{E}[\sum \mathbf{s}\ \mathbf{s}^\top] \leftarrow \mathbf{0}$
$\text{E}[\sum \mathbf{a}\ \mathbf{a}^\top] \leftarrow \mathbf{0}$
for $t = 1$ to T **do**
  $\text{E}[\sum \mathbf{s}\ \mathbf{s}^\top] \leftarrow \text{E}[\sum \mathbf{s}\ \mathbf{s}^\top] + \int_{\mathbf{s}_t} \text{P}(\mathbf{s}_t)\ \mathbf{s}_t\ \mathbf{s}_t^\top\ \partial \mathbf{s}_t$
  $\text{E}[\sum \mathbf{a}\ \mathbf{a}^\top] \leftarrow \text{E}[\sum \mathbf{a}\ \mathbf{a}^\top] + \int_{\mathbf{a}_t} \text{P}(\mathbf{a}_t)\ \mathbf{a}_t\ \mathbf{a}_t^\top\ \partial \mathbf{a}_t$
end for

10.1.2 Latent Information Gradients

The gradients for the latent information setting (Chapter 7) are very similar to those of the previous subsection; they are also differences between empirical feature expectation sums and feature expectations under the maximum causal influence diagram model.

Algorithm 10.3 MaxCausalEnt ID Gradient Calculation

Require: A maximum causal entropy influence diagram and initial parameters, θ.
Ensure: log gradient, $\nabla_\theta \log P(\mathbf{y}||\text{par}(\mathbf{y}))$
Compute $\tilde{E}[\mathcal{F}] \leftarrow \frac{1}{T}\sum_t \text{E}_{\mathbf{X},\mathbf{Y}}[\sum_V \mathcal{F}(V)|\tilde{\mathbf{x}}_t, \tilde{\mathbf{y}}_t]$
Compute $Z(\text{par}(Y)), Z(Y|\text{par}(Y))$ via Algorithm 9.10 for Parameters θ
for all decision nodes Y_i **do**
  Replace Y_i with an uncertainty node with probabilities $\text{P}(y_i|\text{par}(Y_i)) = e^{Z(y_i|\text{par}(Y_i)) - Z(\text{par}(Y_i))}$
end for
$\text{E}[\mathcal{F}] \leftarrow \mathbf{O}^k$
for all V **do**
  for all values $\text{par}(V)'$ **do**
    Compute $\text{P}(\text{par}(V)')$ using $\{\text{P}(y_i|\text{par}(Y_i))\}$
  end for
  $\text{E}[\mathcal{F}] \leftarrow \text{E}[\mathcal{F}] + \text{E}[F(V)]$
end for
$\nabla_\theta \log \text{P}(\mathbf{y}||\text{par}(\mathbf{y})) \leftarrow \tilde{\text{E}}[\mathcal{F}] - \text{E}[\mathcal{F}]$

The added expectation in the maximum causal entropy influence diagram setting preserves the convexity of the optimization since the feature matching constraint remains linear in the causally conditioned probabilities. Standard gradient-based optimization techniques can be employed using the gradient calculated in Algorithm 10.3.

10.1.3 Maximum Causal Entropy Correlated Equilibrium Gradients

In the maximum causal entropy correlated equilibrium setting, optimization of model parameters is also required. In this case, optimization is not employed to match empirical statistics of observed data, but instead to enforce the remaining deviation regret constraints.

From Theorem 8.15 we have the form of the maximum causal entropy correlated equilibrium distribution:

$$\pi_\lambda^{\text{MCECE}}(a_{1:N}^t|s^t) \propto e^{-\left(\sum_{i,a_i^{t\prime}} \lambda_{i,s^t,a_i^t,a_i^{t\prime}} \text{ ExpectDevGain}_i^{\pi^{\text{MCECE}}}\left(a_{1:N}^t,s^t,a_i^{t\prime}\right)\right)+\text{ExpectEnt}\left(a_{1:N}^t,s^t\right)}, \tag{10.4}$$

where $\text{ExpectEnt}(a_{1:N}^t, s^t) \triangleq \mathrm{E}_{a_{1:N}^{t+1},s^{t+1}}\left[\text{ExpectEnt}(a_{1:N}^{t+1}, s^{t+1}) + H(a_{1:N}^{t+1}|s^{t+1})\middle|a_{1:N}^t, s^t\right]$.

The gradient of the log strategy distribution is thus:

$$\frac{\partial}{\partial\lambda_{i,s^t,a_i^t,a_i^{t\prime}}} \log \pi_\lambda^{\text{MCECE}}(a_{1:N}^t|s^t) = \left(\mathrm{P}(a_{1:N}^t|s^t) - 1\right) \text{ ExpectDevGain}_i^\pi(a_{1:N}^t, s^t, a_i^{t\prime}) \tag{10.5}$$

$$\frac{\partial}{\partial\lambda_{i,s^t,a_i^t,a_i^{t\prime}}} \log \pi_\lambda^{\text{MCECE}}(a_{1:N}^t|s^t) = \left(\mathrm{P}(a_{1:N}^t|s^t)\right) \text{ ExpectDevGain}_i^\pi(a_{1:N}^t, s^t, a_i^{t\prime}), \tag{10.6}$$

when each deviation regret constraint is violated (*i.e.*, $\lambda > 0$).

10.2 Convex Optimization Methods

Using the gradient presented in Section 10.1 for each of the maximum causal entropy models, we employ standard gradient-based optimization techniques to fit model parameters from data.

10.2.1 Gradient Descent

The standard, simple gradient-based optimization method is **gradient ascent**. It iteratively takes steps in the direction of the gradient to improve the model parameters according to varying step sizes.

Algorithm 10.4 illustrates the procedure for following the gradient towards the function's optimal point. Standard gradient ascent is often impractically slow for many optimization problems. One standard technique is to replace the decaying step size with **line search**, that finds the optimal step size by performing a binary search along the line in the direction of the gradient.

10.2.2 Stochastic Exponentiated Gradient

When the gradient is computationally expensive to obtain for each training example, computing the entire gradient before taking an optimization step can be extremely slow. Instead, stochastic optimization (Robbins & Monro, 1951) can be employed to take optimization steps after the gradient

Algorithm 10.4 Gradient Ascent calculation

Require: Gradient function $\nabla_\theta F(\mathbf{a}, \mathbf{s}) \rightarrow \mathbb{R}^{|\theta|}$, training data $\{\tilde{a}^{(i)}, \tilde{s}^{(i)}\}$, and learning rate λ.
Ensure: Approximately optimal θ parameters.
1: $\nabla F \leftarrow \mathbf{0}$
2: Initialize θ to random initial values
3: $t \leftarrow 1000$
4: **while** not sufficiently converged **do**
5: **for all** i **do**
6: $\nabla F \leftarrow \nabla F + \nabla_\theta F(\mathbf{a}^{(i)}, \mathbf{s}^{(i)})$
7: **end for**
8: $\theta \leftarrow \theta - \frac{\lambda}{t} \nabla F$
9: $t \leftarrow t + 1$
10: **end while**

for each example or small groups of examples are calculated. The empirical improvements of the stochastic approach have been demonstrated in maximum entropy-based models (Vishwanathan et al., 2006) and other large-scale machine learning tasks. We combine this with the idea of taking exponentiated gradient steps (Kivinen & Warmuth, 1997). This exponentiation techniques has been shown to provide better convergence rates than standard linear gradient update methods.

Algorithm 10.5 Stochastic exponentiated gradient ascent calculation

Require: Gradient function $\nabla_\theta F(\mathbf{a}, \mathbf{s}) \rightarrow \mathbb{R}^{|\theta|}$, training data $\{\tilde{a}^{(i)}, \tilde{s}^{(i)}\}$, and learning rate λ.
Ensure: Approximately optimal θ parameters.
1: Initialize θ to random initial values
2: $t \leftarrow 1000$
3: **while** not sufficiently converged **do**
4: **for all** i in random order **do**
5: $\theta \leftarrow \theta e^{-\frac{\lambda}{t} \nabla_\theta F(\mathbf{a}^{(i)}, \mathbf{s}^{(i)})}$
6: $t \leftarrow t + 1$
7: **end for**
8: **end while**

Algorithm 10.5 illustrates the small differences from gradient ascent that allow stochastic (*i.e.*, online) optimization of the objective function. Often small batches of examples (rather than single examples) are employed to more robustly obtain the gradient step direction of Step 5 of the algorithm.

10.2.3 Subgradient Methods

When the function being optimized is not smooth, but still convex, **sub-gradient methods** can be employed to obtain the optimal parameters. For inequality-constrained optimizations, an alternative to taking optimization steps based on the gradient using the entire set of remaining distribution constraints is to individually apply each violated constraint or a small set of violated constraints using the **sub-gradient** algorithm. This is particularly useful for enforcing inequality constraints.

Given a convex optimization problem of the form:

$$\min_{\theta} f(\theta)$$
$$\text{such that: } \forall_i \; h_i(\theta) \leq 0,$$

the corresponding feasibility optimization can be re-written as an unconstrained optimization (Boyd et al., 2003):

$$\min_{\theta} \max_{i} h_i(\theta). \tag{10.7}$$

The method of subgradient optimization (Shor et al., 1985) can then be applied.

Algorithm 10.6 Sequential constraint, sub-gradient optimization

Require: Inequality constraint functions $\{h\}$, inequality constraint gradient $\{\nabla h\}$, and objective gradient, ∇f, initial parameters, $\hat{\theta}$.
Ensure: Feasible, near-optimal solution to the convex optimization.

```
t ← 1000
while not sufficiently converged or not feasible do
    if a violated constraint exists then
        i ← most violated constraint
        ∇F ← ∇_θ h_i(θ)|_θ̂
    else
        ∇F ← ∇_θ f(θ)|_θ̂
    end if
    θ̂ ← θ̂ − (λ/t)∇F
    t ← t + 1
end while
```

The subgradient optimization of Algorithm 10.6 extends the feasibility optimization of Equation 10.7 to include an objective function. It sequentially finds a violated constraint and takes optimization steps to satisfy that constraint. This process continues until a feasible point is found, at which point an optimization step that improves upon the objective function is taken. We employ this technique to enforce rational inequality constraints in the multi-player game setting.

10.2.4 Other Convex Optimization Methods

A number of more sophisticated methods for convex optimization exist that either employ the Hessian of the objective function, or approximate the Hessian to improve optimization efficiency. Describing each of these methods is beyond the scope of this work. Instead, we treat them as black boxes. The **BFGS** algorithm (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970) is one such technique based on Newton's method of optimization. Interior point methods (Nesterov & Nemirovski, 1987) are also applicable for inequality-constrained optimizations.

These techniques provide better theoretical convergence guarantees than the simpler optimization techniques previously described. However, we employ the simpler methods in practice.

10.3 Considerations for Infinite Horizons

When considering decision problems with infinite horizons, it is often possible for the inference algorithm to be non-convergent for particular values of parameters (as discussed in Section 9.1.3). This complicates simple gradient-based optimization approaches. We present the following three techniques to address this complication.

10.3.1 Projection into Convergent Region

In some settings, constraints on the parameters that enforce convergence can be established. For instance, in the LQR setting, if the $\mathbf{Q}$ matrix is positive-definite, then convergence of the $\mathbf{D}$ matrix over infinite horizons (but ultimately terminating with fixed variance) is guaranteed. This positive definite requirement can be enforced by projecting the $\mathbf{Q}$ matrix using its eigenvalue decomposition and making its negative eigenvalues ϵ.

Convergence-Preserving Gradient Steps

A very similar option is to employ gradient steps that ensure convergent inference. This is difficult in general, but possible for some special cases. For example, exponentiated matrix gradient algorithms (Tsuda et al., 2006) take gradient steps that preserve positive-definiteness. The update step is as follows, for learning rate λ:

$$\mathbf{Q} \leftarrow e^{\log \mathbf{Q} - \lambda_{\mathbf{Q}} \nabla}. \tag{10.8}$$

Here the exponentiation and logarithm operators are matrix operations rather than scalar operators.

10.3.2 Fixed Finite Decision Structures

A second technique to address the infinite horizon and non-convergent inference considerations is to approximate with a fixed horizon, as discussed in Section 9.1.3, or to employ other convergent

inference guaranteeing constraints on the structure of the sequential decision representation. For example, employing discounting naturally forces inference to converge.

10.3.3 Optimization Line-Search and Backtracking

A third technique is to leverage optimization algorithms that employ line search or provide backtracking. Choosing initial parameters that guarantee convergence (*i.e.*, those with very high cost, or large negative reward) is often straight-forward. Gradient-based optimization algorithms that search along the gradient to find the optimal (convergent) point or detect non-convergent inference are then employed. As an example of the latter, in standard gradient-ascent, when a parameter optimization step is taken that leads to non-convergent inference, the algorithm reverts to the last convergent inference parameters and the learning rate is decreased.

10.4 Discussion

In this chapter, we have presented optimization techniques for fitting parameters of the maximum causal entropy models from observed data. The algorithms required for learning are simple applications of common convex optimization techniques. We view this as an advantage; no specialized optimization methods are required and, due to convexity, good optimization convergence guarantees exist. In practice, these simple techniques perform well for training the maximum causal entropy models of this thesis.

Chapter 11

Bayesian Inference with Latent Goals

"People's behavior makes sense if you think about it in terms of their goals, needs, and motives."
— Thomas Mann (Novelist, 1875–1955).

Up until this point of the thesis we have considered the problem setting where intentions—in the form of terminal state constraints on behavior sequence within our maximum causal entropy framework (*i.e.*, terminal utility ϕ from Algorithm 9.1 described in Section 9.1.1)—have been known or assumed. In many problem settings, these intentions, or goals, are not known *a priori* and inferring them given only the partial demonstration of a sequences of states and actions drives a number of important applications. In this chapter, we investigate methods for Bayesian inference of goals and remaining trajectories to unknown goals. Additionally, we provide algorithms with significant efficiency improvements for the setting with deterministic state dynamics.

11.1 Latent Goal Inference

Latent goal inference is the task of predicting the end target, or goal, of a partial sequence of behavior. More formally, a partial sequence of states and actions, $\{\mathbf{s}_{1:t}, \mathbf{a}_{1:t}\}$, is provided. By assumption, when complete, the full sequence will terminate in one state from a set of possible goal states, $G \in \mathcal{G} \subseteq \mathcal{S}$. The probabilistic goal inference problem is to provide a distribution over the possible goal locations.

11.1.1 Goal Inference Approaches

Goal inference can be framed as a probabilistic classification task mapping from the partial trajectory, $\{\mathbf{s}_{1:t}, \mathbf{a}_{1:t}\}$ to a probability distribution over possible goals, $P(\mathcal{G})$. A maximum entropy approach to this problem would define features between the goal variable and partial trajectory

variables, $\mathcal{F}(\{\mathbf{s}_{1:t}, \mathbf{a}_{1:t}\}, G)$. The conditional probability of a goal is then:

$$P(G|\mathbf{s}_{1:t}, \mathbf{a}_{1:t}) \propto e^{\theta^\top \mathcal{F}(\{\mathbf{s}_{1:t}, \mathbf{a}_{1:t}\}, G)} \tag{11.1}$$

This prediction technique provides the worst-case log-loss guarantees of maximum entropy. However, a difficulty of this approach is constructing good features that are a function of the goal state and the partial behavior sequence. For large decision tasks, natural choices of features (*i.e.*, the quadratic interaction terms of partial trajectory states and goal states) can be quite large. This makes fitting the predictive model from a limited amount of observed behavior sequences difficult.

11.1.2 Bayesian Formulation

An alternate approach avoids the potential difficulties of defining a compact set of features relating goals and partial behavior sequences. Using a goal-conditioned distribution as a probabilistic policy, Bayes' rule can be employed to update a belief in which of those goal states the sequence will ultimately terminate. The initial belief is expressed as a prior distribution, $\mathrm{P}(G)$, and the posterior distribution is then:

$$\begin{aligned} \mathrm{P}(G|\mathbf{s}_{1:t}, \mathbf{a}_{1:t}) &= \frac{\mathrm{P}(\mathbf{a}_{1:t}|\mathbf{s}_{1:t}, G)\,\mathrm{P}(G)}{\sum_{G'} \mathrm{P}(\mathbf{a}_{1:t}|\mathbf{s}_{1:t}, G')\,\mathrm{P}(G')} \\ &\propto \left[\prod_t \pi_G(a_t|s_t, G)\right] \mathrm{P}(G), \end{aligned} \tag{11.2}$$

where $\pi_G(a_t|s_t)$ denotes the policy that is constrained to goal state G.

The computational complexity of obtaining this posterior is largely dependent on the complexity of computing each action's conditional probability and the amount of computational re-use possible when calculating conditional probabilities and related quantities.

The naïve approach to latent goal inference computes the posterior probability directly according to Equation 11.2. Algorithm 11.1 illustrates the steps required.

The main bottleneck of this algorithm is that a maximum causal entropy policy must be calculated for each goal state in the goal set. This yields $O(|\mathcal{G}|)$ total policy inference executions (Algorithm 9.1), which dominates the algorithm's run time for a total of $O(|\mathcal{G}|T|\mathcal{S}||\mathcal{A}|)$ or $O(|\mathcal{G}|T(|\mathcal{S}| + |\mathcal{A}|))$ depending on the sparsity of the stochastic transition dynamics (and where T is the number of iterations of the inference algorithm).

11.1.3 Deterministic Dynamics Simplification

In the special case of domains with exclusively deterministic dynamics, we exploit the reversibility of computing the log partition functions to obtain a more efficient algorithm.

Algorithm 11.1 Naïve latent goal inference

Require: MDP $\mathcal{M}_{\text{MDP}}$, trajectory $\mathbf{a}_{1:t}, \mathbf{s}_{1:t}$, set of goals $\mathcal{G} = \{G\}$, prior distribution $\mathrm{P}(G)$.
Ensure: $\{\mathrm{P}(G|\mathbf{a}_{1:t}, \mathbf{s}_{1:t})\}$ posterior distributed according to Equation 11.2.

$Z \leftarrow 0$
for all $G \in \mathcal{G}$ **do**
  Compute $\pi_G(a|s)$ using Algorithm 9.1
  $\mathrm{P}(G|\mathbf{a}_{1:t}, \mathbf{s}_{1:t}) \leftarrow 1$
  for all $\tau = 1$ to t **do**
    $\mathrm{P}(G|\mathbf{a}_{1:t}, \mathbf{s}_{1:t}) \leftarrow \mathrm{P}(G|\mathbf{a}_{1:t}, \mathbf{s}_{1:t})\ \pi_G(a_\tau|s_\tau)$
  end for
  $Z \leftarrow Z + \mathrm{P}(G|\mathbf{a}_{1:t}, \mathbf{s}_{1:t})$
end for
for all $G \in \mathcal{G}$ **do**
  $\mathrm{P}(G|\mathbf{a}_{1:t}, \mathbf{s}_{1:t}) \leftarrow \frac{\mathrm{P}(G|\mathbf{a}_{1:t}, \mathbf{s}_{1:t})}{Z}$
end for

Theorem 11.1. *The posterior distribution of goals can be obtained from the log partition functions, $V_{s_x \to s_y}$, as follows:*

$$\mathrm{P}(G|\boldsymbol{a}_{1:t}, \boldsymbol{s}_{1:t}) \propto \mathrm{P}(G)\ e^{V_{s_t \to G} - V_{s_1 \to G}}. \tag{11.3}$$

The difference in log partition functions, $V_{s_t \to G} - V_{s_1 \to G}$, can be interpreted as the **progress** that the partial sequence has made towards potential goal state, G. Goal state likelihood is maximized by the choice of goal towards which the most progress has been made.

Algorithm 11.2 employs the result of Theorem 11.1 to compute the posterior goal distribution from log partition functions.

Algorithm 11.2 Efficient latent goal inference for deterministic dynamics

Require: MDP $\mathcal{M}_{\text{MDP}}$, trajectory $\mathbf{a}_{1:t}, \mathbf{s}_{1:t}$, set of goals $\mathcal{G} = \{G\}$, prior distribution $\mathrm{P}(G)$.
Ensure: $\mathrm{P}(G|\mathbf{a}_{1:t}, \mathbf{s}_{1:t})$ posterior distributed according to Equation 11.2.

Compute $V_{s_1 \to s_x}$ for all s_x via the forward updates of Algorithm 9.4.
Compute $V_{s_t \to s_x}$ for all s_x via the forward updates of Algorithm 9.4.
for all $G \in \mathcal{G}$ **do**
  $\mathrm{P}(G|\mathbf{a}_{1:t}, \mathbf{s}_{1:t}) \leftarrow \mathrm{P}(G)\ e^{V_{s_t \to G} - V_{s_1 \to G}}$
  $Z \leftarrow Z + \mathrm{P}(G|\mathbf{a}_{1:t}, \mathbf{s}_{1:t})$
end for
for all $G \in \mathcal{G}$ **do**
  $\mathrm{P}(G|\mathbf{a}_{1:t}, \mathbf{s}_{1:t}) \leftarrow \frac{\mathrm{P}(G|\mathbf{a}_{1:t}, \mathbf{s}_{1:t})}{Z}$
end for

Compared to Algorithm 11.1, this approach reduces the required number of policy inferences from $O(|\mathcal{G}|)$ to $O(1)$. This provides a total run time of $O(T(|\mathcal{S}| + |\mathcal{A}|))$. In many settings, this $O(|\mathcal{G}|)$ reduction makes latent goal inference practical for large decision problems.

11.2 Trajectory Inference with Latent Goal State

A related important prediction task is that of inferring a future sequence of state and actions when the goal state is unknown and an initial trajectory, $\mathbf{a}_{1:t}$ and $\mathbf{s}_{1:t}$, has been observed. These tasks build upon the latent goal state inference task by adding trajectory prediction given an inferred goal state distribution.

11.2.1 Bayesian Formulation

Any probabilistic goal inference technique can be employed. However, we will consider the Bayesian latent goal inference perspective. We focus specifically on the problem of calculating the estimated occurrences of some state S_x under the maximum causal entropy distribution with latent goal state:

$$D_{s_x|\mathbf{s}_{1:t},\mathbf{a}_{1:t}} \triangleq \mathrm{E}_{\mathrm{P}(\mathbf{s}_{t+1:T},\mathbf{a}_{t+1:T}|\mathbf{s}_{1:t},\mathbf{a}_{1:t})}\left[\sum_{\tau=t+1}^{T} I(s_x = s_\tau)\middle|\mathbf{s}_{1:t},\mathbf{a}_{1:t}\right]. \tag{11.4}$$

Other related inferences, such as calculating the probability of a continuation trajectory or sub-trajectory can be realized with small modifications.

A simple approach to this inference task is to expand the definition of the expected state visitations (Equation 11.4) and re-expresses it in terms of the latent goal state, G:

$$\begin{aligned} D_{s_x|\mathbf{s}_{1:t},\mathbf{a}_{1:t}} &= \sum_{\mathbf{s}_{t+1:T},\mathbf{a}_{t+1:T}} \left(\sum_{\tau=t+1}^{T} I(s_x = s_\tau)\right) \mathrm{P}(\mathbf{a}_{t+1:T},\mathbf{s}_{t+1:T}|\mathbf{a}_{1:t},\mathbf{s}_{1:t}) \qquad (11.5)\\ &= \sum_{G} \sum_{\mathbf{s}_{t+1:T},\mathbf{a}_{t+1:T}} \left(\sum_{\tau=t+1}^{T} I(s_x = s_\tau)\right) \mathrm{P}(\mathbf{s}_{t+1:T},\mathbf{a}_{t+1:t}|a_t,s_t,G)\,\mathrm{P}(G|\mathbf{s}_{1:t},\mathbf{a}_{1:t}) \\ &= \sum_{G} \sum_{\mathbf{s}_{t+1:T},\mathbf{a}_{t+1:T}} \left(\sum_{\tau=t+1}^{T} I(s_x = s_\tau)\right) \mathrm{P}(\mathbf{s}_{t+1:T},\mathbf{a}_{t+1:t}|a_t,s_t,G)\frac{\mathrm{P}(\mathbf{s}_{1:t},\mathbf{a}_{1:t}|G)\,\mathrm{P}(G)}{\sum_{G'}\mathrm{P}(\mathbf{s}_{1:t},\mathbf{a}_{1:t}|G')\,\mathrm{P}(G')} \end{aligned}$$

The naïve algorithm for latent trajectory inference computes the values of this expansion directly, as shown in Algorithm 11.3.

The total run time is again either $O(|\mathcal{G}|T|\mathcal{S}||\mathcal{A}|)$ or $O(|\mathcal{G}|T(|\mathcal{S}| + |\mathcal{A}|))$ depending on the sparsity of the stochastic transition dynamics.

Algorithm 11.3 Naïve latent trajectory inference

Require: MDP $\mathcal{M}_{\text{MDP}}$, trajectory $\mathbf{a}_{1:t}, \mathbf{s}_{1:t}$, set of goals $\mathcal{G} = \{G\}$, prior distribution $\mathrm{P}(G)$, and state s_x.

Ensure: $D_{s_x|\mathbf{s}_{1:t},\mathbf{a}_{1:t}}$ computed according to Equation 11.4.

$D_{s_x|\mathbf{s}_{1:t},\mathbf{a}_{1:t}} \leftarrow 0$
for all $G \in \mathcal{G}$ **do**
  Compute $\pi_G(a|s)$ using Algorithm 9.6
  Compute $\mathrm{P}(G|\mathbf{a}_{1:t}, \mathbf{s}_{1:t})$ according to Algorithm 11.1
  Compute $D_{s_x,G}$ under $\pi_G(a|s)$ using Algorithm 9.3
  $D_{s_x|\mathbf{s}_{1:t},\mathbf{a}_{1:t}} \leftarrow D_{s_x|\mathbf{s}_{1:t},\mathbf{a}_{1:t}} + D_{s_x,G}\ \mathrm{P}(G|\mathbf{a}_{1:t}, \mathbf{s}_{1:t})$
end for

11.2.2 Deterministic Dynamics Simplification

As in the latent goal inference task, improved efficiency for the latent trajectory inference task can be realized in the special case of deterministic dynamics.

Theorem 11.2. *Action visitation calculations using final reward values are as follows:*

$$\begin{aligned}\phi(G) &= \log P(G|a_{1:t}, s_{1:t}) \\ &= V_{s_t \to G} - V_{s_1 \to G} + \log \mathrm{P}(G),\end{aligned}$$

and are equivalent to the goal-probability-weighted visitation calculations:

$$D_{s_x|\boldsymbol{s}_{1:t},\boldsymbol{a}_{1:t}} = \sum_{\boldsymbol{s}_{t+1:T},\boldsymbol{a}_{t+1:T}} \left(\sum_{\tau=t+1}^{T} I(s_x = s_\tau) \right) \mathrm{P}(\boldsymbol{a}_{t+1:T}, \boldsymbol{s}_{t+1:T}|\boldsymbol{a}_{1:t}, \boldsymbol{s}_{1:t})$$

of Equation 11.4.

Algorithm 11.4 leverages the property of Theorem 11.2 to efficiently calculate state visitation frequencies.

The algorithm requires only one additional inference compared with latent goal inference (Algorithm 11.2), maintaining its overall $O(1)$ inferences and total run time of $O(T(|\mathcal{S}| + |\mathcal{A}|))$.

11.3 Discussion

In this chapter, we have extended the inference procedure for maximum causal entropy to settings where behavior is goal-constrained, but the goal is unknown to the behavior forecaster and must be inferred. Our efficient optimizations for the deterministic dynamics setting generalizes the linear compositionality of Todorov (Todorov, 2009a) to a full Bayesian treatment.

Algorithm 11.4 Efficient latent trajectory inference for deterministic dynamics

Require: MDP $\mathcal{M}_{\text{MDP}}$, trajectory $\mathbf{a}_{1:t}, \mathbf{s}_{1:t}$, set of goals $\mathcal{G} = \{G\}$, prior distribution $\mathrm{P}(G)$, and state s_x.

Ensure: $D_{s_x|\mathbf{s}_{1:t},\mathbf{a}_{1:t}}$ computed according to Equation 11.4

1: Compute $V_{s_t \to s_x}$ for all s_x via the forward updates of Algorithm 9.4.
2: Compute $\{\mathrm{P}(G|\mathbf{a}_{1:t}, \mathbf{s}_{1:t})\}$ via Algorithm 11.2
3: **for all** $G \in \mathcal{G}$ **do**
4: $\quad \phi(G) \leftarrow V_{s_t \to G} - V_{s_1 \to G} + \log \mathrm{P}(G|\mathbf{a}_{1:t}, \mathbf{s}_{1:t})$
5: **end for**
6: Compute $\pi(a|s)$ for final reward $\phi(s)$ using Algorithm 9.1
7: Compute D_{s_x} using Algorithm 9.3, policy $\pi(a|s)$, and initial state s_t

This problem was previously addressed by Baker et al. (2007) under the Boltzmann action value distribution of Section 3.2.5. There are major negative computational efficiency implications for employing that model for goal inference—action values must be computed for each action to all possible goal locations. In contrast, the optimization technique of this chapter for deterministic dynamics requires only a constant number of softmax inferences to obtain posterior goal distributions and posterior trajectories with unknown goal.

Part IV

Applications

Overview of Part IV

Part IV of the thesis employs the models of Part II and the algorithms of Part III to behavior modeling and prediction tasks. We demonstrate that many of the theoretical niceties of the maximum causal entropy framework translate into predictive and task performance improvements over existing methods. Additionally, we show that the algorithms enable these prediction tasks to be performed efficiently.

Chapter 12 applies maximum causal entropy inverse optimal control model to the task of assisting a vehicle's driver by predicting the driver's intended future route in a road network with road characteristics available. We train a maximum causal entropy model using demonstrated global positioning sensor data and evaluate the predictive capabilities.

Chapter 13 applies the maximum causal entropy inverse optimal control model to the task of creating more intelligent robot navigation behavior. Our approach predicts future human movement trajectories and plans robot movements that are more complementary to those predictions.

Lastly, Chapter 14 demonstrates the wider range of applicability of the maximum causal entropy approach on synthetic problems in a multi-player Markov game setting, for predicting behavior in a partial information diagnostic setting, and for predicting the controls of a device in a continuous state and action setting.

Chapter 12

Driver Route Preference Modeling

"Americans will put up with anything provided it doesn't block traffic."
— Dan Rather (Journalist, 1931–).

A common decision task solved by many individuals on a daily basis is that of selecting a driving route to travel from point a to point b in a road network. Many different factors influence this decision task. They range from personal preferences to contextual factors. Global positioning system (GPS) technology allows these decision sequences to be passively observed with no additional interaction burden placed upon the driver. This presents opportunities for:

- Training and evaluating context-dependent decision prediction methods on large-scale sequential decision problems.
- Creating useful applications that benefit drivers based on the ability to accurately predict driver intentions and future behavior.

In this chapter, we apply the maximum causal entropy framework to the problem of modeling driving route preferences and compare with other methods.

12.1 Motivations

Directly communicating our intentions and preferences to a computational system while driving is a difficult (and dangerous) task (Steinfeld et al., 1996). However, a computational system that is aware of that information could be extremely valuable. For example, a navigation application may be able to warn of unexpected accidents or provide more geographically relevant coffee shop recommendations if it knew where a driver were going. Instead of relying on drivers to communicate their intentions, which they often will not or cannot do, we take the opposite perspective; the navigation system itself should learn to predict the intentions and future behaviors of the driver based on past observations and the current situation.

12.2 Understanding Route Preferences

Previous research on predicting the route preferences of drivers found that only 35% of the 2,517 routes taken by 102 different drivers were the "fastest" route, as defined by a popular commercial route planning application (Letchner et al., 2006). Disagreements between route planning software and empirical routes were attributed to contextual factors, like the time of day, and differences in personal preferences between drivers.

We conducted a survey of 21 college students who drive regularly in the Pittsburgh area to help understand the variability of route preference as well as the personal and contextual factors that influence route choice. We presented each participant with four different maps labeled with a familiar start and destination point. In each map we labeled a number of different potentially preferable routes. Participants selected their preferred route from the set of provided routes under 6 different contextual situations for each endpoint pair. The contextual situations we considered were: early weekday morning, morning rush hour, noon on Saturday, evening rush hour, immediately after snow fall, and at night.

Table 12.1: Context-dependent route preference survey results for one pair of endpoints

	Route						
Context	A	B	C	D	E	F	G
Early morning	6	6	4	1	2	2	0
Morning rush hour	8	4	5	0	1	2	1
Saturday noon	7	5	4	0	1	2	2
Evening rush hour	8	4	5	0	0	3	1
After snow	7	4	4	3	2	1	0
Midnight	6	4	4	2	1	4	0

The routing problem most familiar to our participants had the most variance of preference. The number of participants who most preferred each route under each of the contextual situations for this particular routing problem is shown in Table 12.1. Of the 7 available routes to choose from (A-G) under 6 different contexts, the route with highest agreement was only preferred by 8 people (38%). In addition to route choice being dependent on personal preferences, route choice was often context-dependent. Of the 21 participants, only 6 had route choices that were context-invariant. 11 participants used two different routes depending on the context, and 4 participants employed three different context-dependent routes.

Our participants were not the only ones in disagreement over the best route for this particular endpoint pair. We generated route recommendations from four major commercial mapping and direction services for the same endpoints to compare against. The resulting route recommendations also demonstrated significant variation, though all services are context- and preference-independent. Google Maps (Google 2008) and Microsoft's MapPoint (MapPoint 2008) both chose

route E, while MapQuest (MapQuest 2008) generated route D, and Yahoo! Maps (Yahoo 2008) provided route A.

Participants additionally provided their preference towards different driving situations on a five-point Likert scale (see Table 12.2). The scale was defined as: (1) very strong dislike, avoid at all costs; (2) dislike, sometimes avoid; (3) don't care, doesn't affect route choice; (4) like, sometimes prefer; (5) very strong like, always prefer.

Table 12.2: Situational preference survey results

	Preference				
Situation	1	2	3	4	5
Interstate/highway	0	0	3	14	4
Excess of speed limit	1	4	5	10	1
Exploring unfamiliar area	1	8	4	7	1
Stuck behind slow driver	8	10	3	0	0
Longer routes with no stops	0	4	3	13	1

While some situations, like driving on the interstate are disliked by no one and being stuck behind a slow driver are preferred by no one, other situations have a wide range of preference with only a small number of people expressing indifference. For example, the number of drivers who prefer routes that explore unfamiliar areas is roughly the same as the number of drivers with the opposite preference, and while the majority prefer to drive in excess of the speed limit and take longer routes with no stops, there were a number of others with differing preferences. We expect that a population covering all ranges of age and driving ability will possess an even larger preference variance than the more homogeneous participants in our surveys.

The results of our formative research strongly suggest that drivers' choices of routes vary greatly and are highly dependent on personal preferences and contextual factors.

12.3 PROCAB: Context-Aware Behavior Modeling

The variability of route preference from person to person and from situation to situation makes perfectly predicting every route choice for a group of drivers extremely difficult. We adopt the more modest goal of developing a *probabilistic model* that assigns as much probability as possible to the routes the drivers prefer. Some of the variability from personal preference and situation is explained by incorporating contextual variables within our probabilistic model. The remaining variability in the probabilistic model stems from influences on route choices that are unobserved by our model.

Many different assumptions and frameworks can be employed to probabilistically model the relationships between contextual variables and sequences of actions. The probabilistic reasoning

from observed context-aware behavior (PROCAB) approach that we employ is based on three principled techniques previously introduced in this thesis:

- Representing driving route decisions as sequential actions in a **parametric-reward Markov decision process** (PRMDP) with parametric costs (Definition 2.10).
- Using **inverse optimal control** (Section 3.2.1) to recover cost weights for the PRMDP that explain observed behavior.
- Employing the **principle of maximum entropy** (Section 5.1) and its causal extension (Section 5.2) to find the unique set of cost weights that have the least commitment.

The resulting probabilistic model of behavior is context-dependent, compact, and efficient to learn and reason about. It can be viewed as a special deterministic case of the maximum causal entropy approach to inverse optimal control. In this section, we describe the formulation of the route selection task as a Markov decision process. We then explain how the resulting model is employed to efficiently reason about context-dependent behavior.

12.3.1 Representing Routes Using a Markov Decision Process

Markov decision processes (MDPs) (Puterman, 1994) provide a natural framework for representing sequential decision making tasks, such as route planning. The agent takes a sequence of *actions* ($a \in A$), which transition the agent between *states* ($s \in S$) and incur an action-based *cost*[1] ($c(a) \in \mathbb{R}$). We employ the PRMDP variant where costs of actions are parameterized (Definition 2.10). A simple deterministic PRMDP with 8 states and 20 actions representing a small grid-like road network is shown in Figure 12.1.

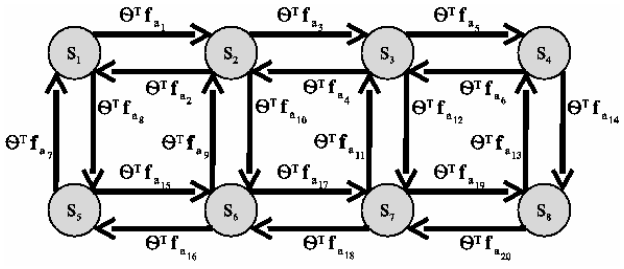

Figure 12.1: A simple Markov Decision Process with action costs.

In the optimal control problem, the agent is trying to minimize the sum of costs while reaching some destination. We call the sequence of actions a *path*, ζ. For MDPs with parametric costs, a set of *features* ($\mathbf{f}_a \in \mathbb{R}^k$) characterize each action, and the cost of the action is a linear function of

[1]The negation of costs, *rewards*, are more common in the MDP literature, but less intuitive for our application.

these features parameterized by a *cost weight* vector ($\theta \in \mathbb{R}^k$). Path features, $\mathbf{f}_\zeta$, are the sum of the features of actions in the path: $\sum_{a \in \zeta} \mathbf{f}_a$. The path cost is the sum of action costs (Figure 12.1), or, equivalently, the cost weight applied to the path features.

$$\text{cost}(\zeta|\theta) = \sum_{a \in \zeta} \theta^\top \mathbf{f}_a = \theta^\top \mathbf{f}_\zeta$$

The advantage of the MDP approach is that the cost weight is a compact set of variables representing the reasons for preferring different behavior, and if it is assumed that the agent acts sensibly with regard to incurred costs, the approach generalizes to previously unseen situations.

12.4 Taxi Driver Route Preference Data

Now that we have described our model for probabilistically reasoning from observed context-aware behavior, we will describe the data we collected to evaluate this model. We recruited 25 Yellow Cab taxi drivers from whom we collected GPS data. Their experience as taxi drivers in Pittsburgh ranged from 1 month to 40 years. The average and median were 12.9 years and 9 years respectively. All participants reported living in the area for at least 15 years.

12.4.1 Collected Position Data

We collected location traces from our study participants over a three month period using global positioning system (GPS) devices that log locations over time. Each participant was provided one of these devices[2], which records a reading roughly every 6 to 10 seconds while in motion. The data collection yielded a dataset of over 100,000 miles of travel collected from over 3,000 hours of driving. It covers a large area surrounding our city (Figure 12.2). Note that no map is being overlaid in this figure. Repeated travel over the same roads leaves the appearance of the road network itself.

12.4.2 Road Network Representation

The deterministic action-state representation of the corresponding road network contains over 300,000 states (*i.e.*, road segments) and over 900,000 actions (*i.e.*, available transitions between road segments). There are characteristics describing the speed categorization, functionality categorization, and lane categorization of roads. Additionally we can use geographic information to obtain road lengths and turn types at each intersection (*e.g.*, straight, right, left).

[2]In a few cases where two participants who shared a taxi also shared a GPS logging device

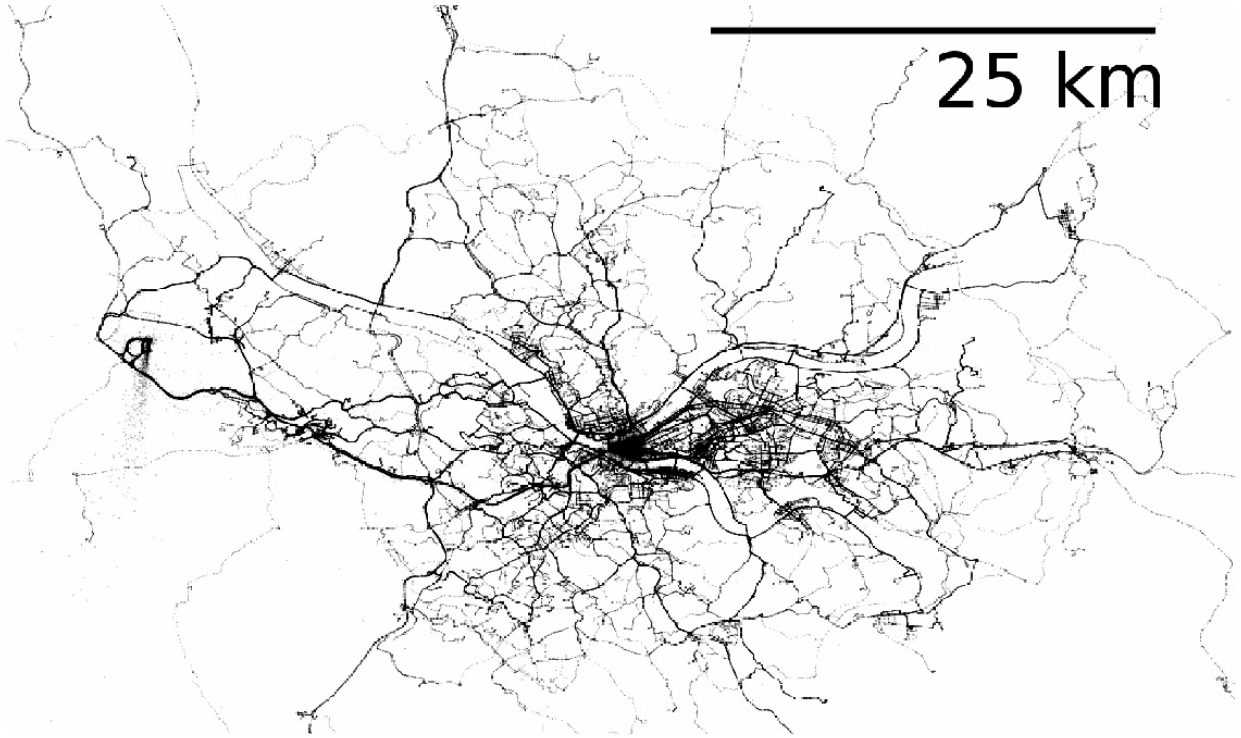

Figure 12.2: The collected GPS datapoints

12.4.3 Fitting to the Road Network and Segmenting

To address noise in the GPS data, we fit it to the road network using a particle filter (Thrun et al., 2005). A particle filter simulates a large number of vehicles traversing over the road network, focusing its attention on particles that best match the GPS readings. A motion model is employed to simulate the movement of the vehicle and an observation model is employed to express the relationship between the true location of the vehicle and the GPS reading of the vehicle. We use a motion model based on the empirical distribution of changes in speed and a Laplace distribution for our observation model.

Once fitted to the road network, we segmented our GPS traces into distinct trips. Our segmentation is based on time-thresholds. Position readings with a small velocity for a period of time are considered to be at the end of one trip and the beginning of a new trip. We note that this problem is particularly difficult for taxi driver data, because these drivers may often stop only long enough to let out a passenger and this can be difficult to distinguish from stopping at a long stoplight. To address same of the potential noise, we discard trips that are too short, too noisy, and too cyclic.

12.5 Modeling Route Preferences

In this section, we describe the features that we employ in the maximum causal entropy prediction approach. These features serve to model the utilities of different road segments in our predictive model.

12.5.1 Feature Sets and Context-Awareness

As mentioned, we have characteristics of the road network that describe speed, functionality, lanes, and turns. These combine to form path-level features that describe a path as *numbers of different turn types along the path* and road mileage at different:

- speed categories,
- functionality categories, and
- numbers of lanes.

To form these path features, we combine the comprising road segments' features for each of their characteristics weighted by the road segment length or intersection transition count.

Table 12.3: Example feature counts for a driver's demonstrated route(s)

Feature	Value
Highway	3.3 miles
Major Streets	2.0 miles
Local Streets	0.3 miles
Above 55mph	4.0 miles
35-54mph	1.1 miles
25-34 mph	0.5 miles
Below 24mph	0 miles
3+ Lanes	0.5 miles
2 Lanes	3.3 miles
1 Lane	1.8 miles

Feature	Value
Hard left turn	1
Soft left turn	3
Soft right turn	5
Hard right turn	0
No turn	25
U-turn	0

The PROCAB approach finds the cost weight for different features so that the model's feature counts will match (in expectation) those demonstrated by a driver (*e.g.*, as shown in Table 12.3). when planning for the same starting point and destination. Additionally, unobserved features may make certain road segments more or less desirable. To model these unobservable features, we add unique features associated with each road segment to our model. This allows the cost of each road segment to vary independently.

We conducted a survey of our taxi drivers to help identify the main contextual factors that impact their route choices. The perceived influences (with average response) on a 5-point Likert Scale ranging from "no influence" (1) to "strong influence" (5) are: *Accidents* (4.50), *Construction* (4.42), *Time of Day* (4.31), *Sporting Events* (4.27), and *Weather* (3.62). We model some of these influences by adding real-time features to our road segments for *Accident*, *Congestion*, and *Closures* (Road and Lane) according to traffic data collected every 15 minutes from Yahoo's traffic service.

We incorporate sensitivity to time of day and day of week by adding duplicate features that are only active during certain times of day and days of week. We use *morning rush hour*, *day time*, *evening rush hour*, *evening*, and *night* as our time of day categories, and *weekday* and *weekend* as day of week categories. Using these features, the model tries to not only match the total number of *e.g.*, interstate miles, but it also tries to match the right number of interstate miles under each time of day category. For example, if our taxi drivers try to avoid the interstate during rush hour, the PROCAB model assigns a higher weight to the joint interstate and rush hour feature. It is then less likely to prefer routes on the interstate during rush hour given that higher weight.

Matching all possible contextual feature counts highly constrains the model's predictions. In fact, given enough contextual features, the model may exactly predict the actual demonstrated behavior and *overfit* to the training data. We avoid this problem using *regularization* (for theoretical justification, see Section 5.1.4), a technique that relaxes the feature matching constraint by introducing a penalty term ($-\sum_i \lambda_i \theta_i^2$) to the optimization. This penalty prevents cost weights corresponding to highly specialized features from becoming large enough to force the model to perfectly fit observed behavior.

12.5.2 Learned Cost Weights

We learn the cost weights that best explain a set of demonstrated routes using Algorithm 9.6. For improved efficiency, we also consider only the set of paths residing within some sub-portion of the road network. The sub-portion is carefully chosen to minimize the approximation loss while maximizing algorithm speed. Using this approach, we can obtain cost weights for each driver or a collective cost weight for a group of drivers. In this work, we group the routes gathered from all 25 taxi drivers together and learn a single cost weight using a training set of 80% of those routes.

Figure 12.3 shows how road type and speed categorization influence the road's cost in our learned model. The generally monotonic relation between learned costs and these categories of road type match what we might assume about taxi drivers' preferences.

12.6 Navigation Applications and Evaluation

We now focus on the applications that our route preference model enables. We evaluate our model on a number of prediction tasks needed for those applications. We compare the PROCAB model's

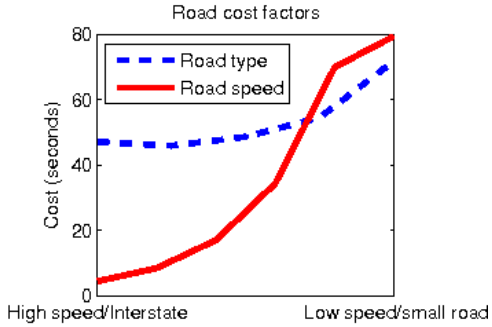

Figure 12.3: Speed categorization and road type cost factors normalized to seconds assuming 65mph driving on fastest and largest roads

performance with other state of the art methods on these tasks using the remaining 20% of the taxi route dataset to evaluate.

12.6.1 Turn Prediction

We first consider the problem of predicting the action at the next intersection given a known final destination. This problem is important for applications such as automatic turn signaling and fuel consumption optimization when the destination is either specified by the driver or can be inferred from his or her normal driving habits. We compare PROCAB's ability to predict 55,725 decisions at intersections with multiple options[3] to approaches based on Markov models.

We measure the accuracy of each model's predictions (*i.e.*, percentage of predictions that are correct) and the average log likelihood, $\frac{1}{\#decisions} \sum_{\text{decision } d} \log \mathrm{P}_{\text{model}}(d)$ of the actual actions in the model. This latter metric evaluates the model's ability to probabilistically estimate the distribution of decisions. Values closer to zero are closer to perfectly predicting the data. As a baseline, guessing uniformly at random (without U-turns) yields an accuracy of 46.4% and a log likelihood of -0.781 for our dataset.

Markov Models for Turn Prediction

We implemented two Markov models (introduced in Section 3.1.1) for turn prediction. A Markov model predicts the next action (or road segment) given some observed conditioning variables. Its predictions are obtained by considering a set of previously observed decisions at an intersection that match the conditioning variables. The most common action from that set can be predicted or the probability of each action can be set in proportion to its frequency in the set[4]. For example,

[3]Previous Markov model evaluations include "intersections" with only one available decision, which comprise 95% (Simmons et al., 2006) and 28% (Krumm, 2008) of decisions depending on the road network representation.

[4]Some chance of selecting previously not taken actions is enabled by *smoothing* the distribution.

Liao et al. (2007b) employ the destination and mode of transport for modeling multi-modal transportation routines. We evaluate the Markov models employed by Krumm (2008), which conditions on the previous *K* traversed road segments, and Simmons et al. (2006), which conditions on the route destination and current road segment being traversed.

Table 12.4: *K*-order Markov model performance

	Non-reducing		Reducing	
History	**Accuracy**	**Likelihood**	**Accuracy**	**Likelihood**
1 edge	**85.8%**	−0.322	85.8%	−0.322
2 edge	85.6%	**−0.321**	86.0%	**−0.319**
5 edge	84.8%	−0.330	**86.2%**	−0.321
10 edge	83.4%	−0.347	86.1%	−0.328
20 edge	81.5%	−0.367	85.7%	−0.337
100 edge	79.3%	−0.399	85.1%	−0.356

Evaluation results of a *K*-order Markov model based on road segment histories are shown in Table 12.4. We consider two variants of the model. For the *non-reducing* variant, if there are no previously observed decisions that match the K-sized history, a random guess is made. In the *Reducing* variant, instead of guessing randomly when no matching histories are found, the model tries to match a smaller history size until some matching examples are found. If no matching histories exist for $K = 1$ (*i.e.*, no previous experience with this decision), a random guess from the possible next actions is then made.

In the non-reducing model we notice a performance degradation as the history size, K, increases. The reduction strategy for history matching helps to greatly diminish the degradation of having a larger history, but we still find a small performance degradation as the history size increases beyond 5 edges.

Table 12.5: Destination Markov model performance

Destinations	**Accuracy**	**Likelihood**
1x1 grid	**85.8%**	−0.322
2x2 grid	85.1%	**−0.319**
5x5 grid	84.1%	−0.327
10x10 grid	83.5%	−0.327
40x40 grid	78.6%	−0.365
100x100 grid	73.1%	−0.416
2000x2000 grid	59.3%	−0.551

We present the evaluation of the Markov model conditioned on a known destination in Table

12.5. We approximate each unique destination with a grid of destination cells, starting from a single cell (1x1) covering the entire map, all the way up to 4 million cells (2000x2000). As the grid dimensions grow to infinity, the approach treats each destination as unique. We find empirically that using finer grid cells actually degrades accuracy, and knowing the destination provides no advantage over the history-based Markov model for this task.

Both of these results show the inherent problem of *data sparsity* for directed graphical models in these types of domains. With an infinite amount of previously observed data available to construct a model, having more information (*i.e.*, a larger history or a finer grid resolution) can never degrade the model's performance. However, with finite amounts of data there will often be few or no examples with matching conditional variables, providing a poor estimate of the true conditional distribution, which leads to lower performance in applications of the model.

PROCAB Turn Prediction

The PROCAB model predicts turns by reasoning about paths to the destination. Each path has some probability within the model, and many different paths share the same first actions. An action's probability is obtained by summing up all path probabilities that start with that action. The PROCAB model provides a compact representation over destination, context, and action sequences, so the exact destination and rich contextual information can be incorporated without leading to data sparsity degradations like the Markov model.

Table 12.6: Baseline and PROCAB turn prediction performance

	Accuracy	**Likelihood**
Random guess	46.4%	−0.781
Best history Markov model	86.2%	−0.319
Best destination Markov model	85.8%	−0.319
PROCAB (no context)	91.0%	−0.240
PROCAB (context)	**93.2%**	**−0.201**

We summarize the best comparison models' results for turn prediction and present the PROCAB turn prediction performance in Table 12.6. The PROCAB approach provides a large improvement over the other models in both prediction accuracy and log likelihood. Additionally, incorporating time of day, day of week, and traffic report contextual information provides improved performance. We include this additional contextual information in the remainder of our experiments.

12.6.2 Route Prediction

We now focus on route prediction, where the origin and destination of a route are known, but the route between the two is not and must be predicted. Two important applications of this problem are route recommendation, where a driver requests a desirable route connecting two specified points, and unanticipated hazard warning, where an application can predict the driver will encounter some hazard he is unaware of and provide a warning beforehand so that the hazard can be avoided.

We evaluate the prediction quality based on the amount of matching distance between the predicted route and the actual route, and the percentage of predictions that match. We consider all routes that share 90% of distance as matches. This final measure ignores minor route differences, like those caused by noise in GPS data. We evaluate a previously described Markov model, a model based on estimated travel time, and our PROCAB model.

Markov Model Route Planning

We employ route planning using the previously described destination-conditioned Markov model (Simmons et al., 2006). The model recommends the most probable route satisfying origin and destination constraints.

The results (Table 12.7, *Markov*) are fairly uniform regardless of the number of grid cells employed, though there is a subtle degradation with more grid cells.

Travel Time-Based Planning

A number of approaches for vehicle route recommendation are based on estimating the travel time of each route and recommending the fastest route. Commercial route recommendation systems, for example, try to provide the fastest route. The Cartel Project (Hull et al., 2006) works to better estimate the travel times of different road segments in near real-time using fleets of instrumented vehicles. One approach to route prediction is to assume the driver will also try to take this most expedient route.

We use the distance and speed categorization of each road segment to estimate travel times, and then provide route predictions using the fastest route between origin and destination. The results (Table 12.7, *travel time*) show a large improvement over the Markov model. We believe this is due to data sparsity in the Markov model and the travel time model's ability to generalize to previously unseen situations (*e.g.*, new origin-destination pairs).

Inverse Optimal Control and PROCAB

Our view of route prediction and recommendation is fundamentally different than those based solely on travel time estimates. Earlier research (Letchner et al., 2006) and our own study have shown that there is a great deal of variability in route preference between drivers. Rather than assume drivers are trying to optimize one particular metric and more accurately estimate that metric

in the road network, we implicitly learn the metric that the driver is actually trying to optimize in practice. This allows other factors on route choice, such as fuel efficiency, safety, reduced stress, and familiarity, to be modeled. While the model does not explicitly understand that one route is more stressful than another, it does learn to imitate a driver's avoidance of more stressful routes and features associated with those routes.

We first evaluate two other IRL models. The first is Maximum Margin Planning (MMP) Ratliff et al. (2006), which is a model capable of predicting new paths, but incapable of density estimation (i.e., computing the probability of some demonstrated path). The second model is an action-based distribution model (Action) that has been employed for Bayesian IRL Ramachandran & Amir (2007) and hybrid IRL Neu & Szepesvári (2007). The choice of action in any particular state is assumed to be distributed according to the future expected reward of the best policy after taking the action, $Q^*(S, a)$. In our setting, this value is simply the optimal path cost to the goal after taking a particular action.

Table 12.7: Evaluation results for Markov model with various grid sizes, time-based model, the PROCAB model, and other inverse optimal control approaches

Model	**Dist. Match**	**90% Match**
Markov (1x1)	62.4%	30.1%
Markov (3x3)	62.5%	30.1%
Markov (5x5)	62.5%	29.9%
Markov (10x10)	62.4%	29.6%
Markov (30x30)	62.2%	29.4%
Travel Time	72.5%	44.0%
Max Margin	75.3%	46.6%
Action	77.3%	50.4%
Action (costs)	77.7%	50.8%
PROCAB	**82.6%**	**61.0%**

Inverse optimal control approaches and the PROCAB model specifically provide increased performance over both the Markov model and the model based on travel time estimates, as shown in Table 12.7. This is because the inverse optimal control approaches are able to better estimate the utilities of the road segments than the feature-based travel time estimates. Additionally, the PROCAB model outperforms the other inverse optimal control approaches. One explanation is that its capability to address the inherent sub-optimality of demonstrated taxi trajectories is better than the other approaches.

12.6.3 Destination Prediction

Finally, we evaluate models for destination prediction. In situations where a driver has not entered her destination into an in-car navigation system, accurately predicting her destination would be useful for proactively providing an alternative route. Given the difficulty that users have in entering destinations (Steinfeld et al., 1996), destination prediction can be particularly useful. It can also be used in conjunction with our model's route prediction and turn prediction to deal with settings where the destination is unknown.

In this setting, a partial route of the driver is observed and the final destination of the driver is predicted. This application is especially difficult given our set of drivers, who visit a much wider range of destinations than typical drivers. We compare PROCAB's ability to predict destination to two other models, in settings where the set of possible destinations is not fully known beforehand.

We evaluate our models using 1881 withheld routes and allow our model to observe various amounts of the route from 10% to 90%. The model is provided no knowledge of how much of the trip has been completed. Each model provides a single point estimate for the location of the intended destination and we evaluate the distance between the true destination and this predicted destination in kilometers.

Bayes' Rule

As discussed in Chapter 11, we employ Bayes' rule (Equation 12.1) and probabilistic route preference models that predict route (B) given destination (A) can be employed along with a prior on destinations, $\mathrm{P}(A)$, to obtain a probability distribution over destinations, $\mathrm{P}(A|B)$[5]:

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\,\mathrm{P}(A)}{\sum_{A'} \mathrm{P}(B|A')\,\mathrm{P}(A')} \propto \mathrm{P}(B|A)\,\mathrm{P}(A). \tag{12.1}$$

Markov Model Destination Prediction

We first evaluate a destination-conditioned Markov model for predicting destination region in a grid. The model employs Bayes' rule to obtain a distribution over cells for the destination based on the observed partial route. The prior distribution over grid cells is obtained from the empirical distribution of destinations in the training dataset. The center point of the most probable grid cell is used as a destination location estimate.

Evaluation results for various grid cell sizes are shown in Table 12.8. We find that as more of the driver's route is observed, the destination is more accurately predicted. As with the other applications of this model, we note that having the most grid cells does not provide the best model due to data sparsity issues.

[5]Since the denominator is constant with respect to A, the probability is often expressed as being proportionate to the numerator.

Table 12.8: Prediction error of Markov, Predestination, and PROCAB models in kilometers

	Percentage of trip observed					
Model	**10%**	**20%**	**40%**	**60%**	**80%**	**90%**
Markov model 5x5	9.61	9.24	8.75	8.65	8.34	8.17
Markov model 10x10	9.31	8.50	7.91	7.58	7.09	6.74
Markov model 20x20	9.69	8.99	8.23	7.66	6.94	6.40
Markov model 30x30	10.5	9.94	9.14	8.66	7.95	7.59
Predestination 40x40 MAP	7.98	7.74	6.10	5.42	4.59	4.24
Predestination 40x40 Mean	7.74	7.53	5.77	4.83	4.12	3.79
Predestination 80x80 MAP	11.02	7.26	5.17	4.86	4.21	3.88
Predestination 80x80 Mean	8.69	7.27	5.28	**4.46**	**3.95**	3.69
PROCAB MAP	11.18	8.63	6.63	5.44	3.99	**3.13**
PROCAB Mean	**6.70**	**5.81**	**5.01**	4.70	3.98	3.32

Predestination

The Predestination system (Krumm & Horvitz, 2006) grids the world into a number of cells and uses the observation that the partially traveled route is usually an *efficient* path through the grid to the final destination. Using Bayes' rule, destinations that are opposite of the direction of travel have much lower probability than destinations for which the partially traveled route is efficient. Our implementation employs a prior distribution over destination grid cells conditioned on the starting location rather than the more detailed original Predestination prior. We consider two variants of prediction with Predestination. One predicts the center of the most probable cell (*i.e.*, the *Maximum a Posteriori* or *MAP* estimate). The other, *Mean*, predicts the probabilistic average over cell center beliefs. We find that Predestination shows significant improvement (Table 12.8) over the Markov model's performance.

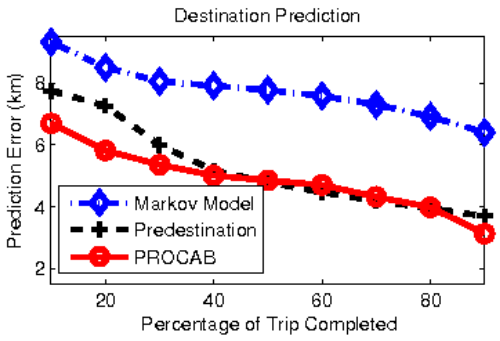

Figure 12.4: The best Markov model, Predestination, and PROCAB prediction errors

PROCAB Destination Prediction

In the PROCAB model, destination prediction is also an application of Bayes' rule. Consider a *partial path*, $\zeta_{A\to B}$ from point A to point B. The destination probability is then:

$$\begin{aligned}\mathrm{P}(\text{dest}|\zeta_{A\to B},\theta) &\propto \mathrm{P}(\zeta_{A\to B}|\text{dest},\theta)\,\mathrm{P}(\text{dest}) \\ &\propto \frac{\sum_{\zeta_{B\to\text{dest}}} e^{\text{reward}(\zeta|\theta)}}{\sum_{\zeta_{A\to\text{dest}}} e^{\text{reward}(\zeta|\theta)}}\,\mathrm{P}(\text{dest}) \qquad (12.2)\end{aligned}$$

We use the same prior that depends on the starting point (A) that we employed in our Predestination implementation. The posterior probabilities are efficiently computed by taking the sums over paths from points A and B to each possible destination (Equation 12.2) using the forward pass of Algorithm 1.

Table 12.8 and Figure 12.4 show the accuracy of PROCAB destination prediction compared with other models. Using averaging is much more beneficial for the PROCAB model, likely because it is more sensitive to particular road segments than models based on grid cells. We find that the PROCAB model performs comparably to the Predestination model given a large percentage of observed trip, and better than the Predestination model when less of the trip is observed. PROCAB's abilities to learn non-travel time desirability metrics, reason about road segments rather than grid cells, and incorporate additional contextual information may each contribute to this improvement.

12.7 Discussion

In this chapter, we have applied the maximum causal entropy approach to prediction tasks in personal navigation. We have demonstrated state-of-the-art performance on turn prediction, route prediction, and destination prediction tasks using large-scale data collected from taxi drivers. One of the main advantages of our approach—personalization to individual drivers driving preferences—is not well-explored by this approach, since taxi drivers are typically optimizing for travel-time rather than *e.g.*, avoiding certain types of driving for safety reasons. We expect an even larger performance improvement over more naïve approaches on prediction tasks for non-professional drivers.

Chapter 13

Pedestrian Motion Prediction

"Real men usually voluntarily avoid robots"

— Tadokoro et al. (1995)

In this chapter, we apply the maximum causal entropy framework to the problem of predicting the future movements of a pedestrian within an indoor environment. Accurate predictions in this task are important for planning intelligent robot movements that appropriately balance the desire to have the robot efficiently reach its destination with the desire not to hinder peoples' movements in the environment.

13.1 Motivations

Determining appropriate robotic actions in environments with moving people is a well-studied (Tadokoro et al., 1995; Bennewitz et al., 2002; Foka & Trahanias, 2002), but often difficult task due to the uncertainty of each person's future behavior. Robots should certainly never collide with people (Petti & Fraichard, 2005), but avoiding collisions alone is often unsatisfactory because the disruption of almost colliding can be burdensome to people and sub-optimal for robots. Instead, robots should predict the future locations of people and plan routes that will avoid such hindrances (*i.e.*, situations where the person's natural behavior is disrupted due to a robot's proximity) while still efficiently achieving the robot's objectives. For example, given the origins and target destinations of the robot and person in Figure 13.1, the robot's hindrance-minimizing trajectory would take the longer way around the center obstacle (a table), leaving a clear path for the pedestrian.

One common approach for predicting trajectories is to project the prediction step of a tracking filter (Madhavan & Schlenoff, 2003; Schlenoff et al., 2004; Mertz, 2004) forward over time. For example, a Kalman filter's (Kalman & Bucy, 1962) future positions are distributed according to a Gaussian distribution with growing uncertainty and, unfortunately, often high probability for physically impossible locations (*e.g.*, behind walls, within obstacles). Particle filters (Thrun et al., 2005) can incorporate more sophisticated constraints and non-Gaussian distributions, but degrade

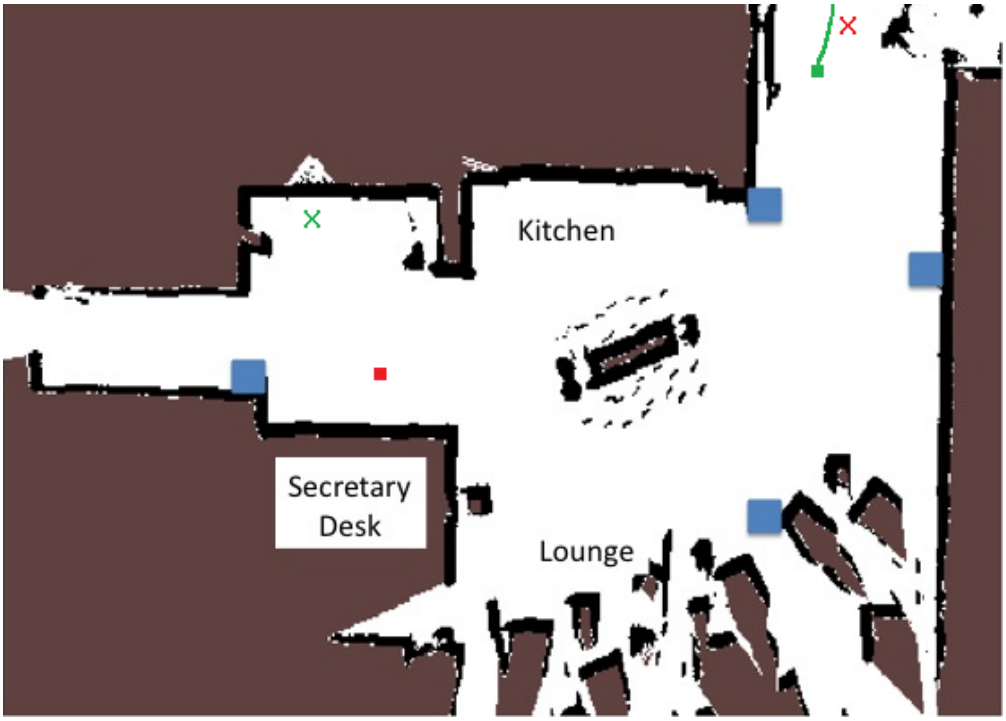

Figure 13.1: A hindrance-sensitive robot path planning problem in our experimental environment containing a person (green square) in the upper right with a previous trajectory (green line) and intended destination (green X) near a doorway, and a robot (red square) near the secretary desk with its intended destination (red X) near the person's starting location. Hindrances are likely if the person and robot both take the distance-minimizing path to their intended destinations. Laser scanners are denoted with blue boxes.

into random walks of feasible rather than purposeful motion over large time horizons. Closer to our research are approaches that directly model the policy (Galata et al., 2001). These approaches assume that previously observed trajectories capture all purposeful behavior, and the only uncertainty involves determining to which previously observed class of trajectories the current behavior belongs. Models based on mixtures of trajectories and conditioned action distribution modeling (using hidden Markov models) have been employed (Vasquez Govea et al., 2005). This approach often suffers from over-fitting to the particular training trajectories and context of those trajectories. When changes to the environment occur (*e.g.*, rearrangement of the furniture), the model will confidently predict incorrect trajectories *through* obstacles.

We assume that people behave similarly to planners, *i.e.*, they efficiently move to reach destinations. In traditional planning, given a *cost function* mapping environment features to costs, the optimal trajectory is easily obtained for any endpoints in any environment described using those features. Our approach learns the cost function that best explains previously observed trajectories. Unfortunately, traditional planning is *prescriptive* rather than *predictive*—the sub-optimality typi-

Figure 13.2: Images of the kitchen area (left), secretary desk area (center), and lounge area (right) of our experimental environment.

cally present in observed data is inexplicable to a planner. We employ the principle of **maximum causal entropy** developed in this thesis to address the lack of decision uncertainty. Specifically, we employ the **maximum causal entropy inverse optimal control** technique of Chapter 6. This approach yields a *soft-maximum* version of Markov decision processes (MDP) that accounts for decision uncertainty. As we show, this soft-max MDP model supports efficient algorithms for learning the cost function that best explains previous behavior, and for predicting a person's future positions.

Importantly, the featured-based cost function that we employ enables generalization. Specifically, the cost function is a linear combination of a given set of features computed from the environment (*e.g.*, obstacles and filters applied to obstacles). Once trained, the cost function applies to any configuration of these features. Therefore if obstacles in the environment move, the environment otherwise changes, or we consider an entirely different environment, our model generalizes to this new setting. We consider this improved generalization to be a major benefit of our approach over previous techniques.

Predictions of pedestrian trajectories can be naturally employed by a planner with time-dependent costs so that potential hindrances are penalized. Unfortunately, the increased dimensionality of the planning problem can be prohibitive. Instead, we present a simple, incremental "constraint generation" planning approach that enables real-time performance. This approach initially employs a cost map that ignores the predictions of people's future locations. It then iteratively plans the robot's trajectory in the cost map, simulates the person's trajectory, and adds "cost" to the cost map based on the probability of hindrance at each location. The time-independent cost function that this procedure produces accounts for the time-varying predictions, and ultimately yields a high quality, hindrance-free robot trajectory, while requiring much less computation than a time-based planner.

We evaluate the quality of our combined prediction and planning system on the trajectories of people in a lab environment using the opposing objectives of maximizing the robot's efficiency in reaching its intended destination and minimizing robot-person hindrances. An inherent trade-off

Algorithm 13.1 Incorporating predictive pedestrian models via predictive planning

PredictivePlanning($\sigma > 0,\ \alpha > 0$, $\{D_{s,t}\}$, D_{thresh})
Initialize cost map to prior navigational costs $c_0(s)$.
for $t = 0, \ldots, T$ **do**
    Plan under the current cost map.
    Simulate the plan forward to find points of probable interference with the pedestrian $\{(s_i)\}_{i=1}^{K_t}$ where $D_{s,t} > D_{\text{thresh}}$.
    If $K = 0$ then break.
    Add cost to those points
        $c_{t+1}(s) = c_t(s) + \alpha \sum_{i=1}^{K_t} e^{-\frac{1}{2\sigma^2}\|s-s_i\|^2}$.
end for
Return the plan through the final cost map.

between these two criteria exists in planning appropriate behavior. We show that for nearly any chosen trade-off, our prediction model is better for making decisions than an alternate approach.

13.2 Planning with Pedestrian Predictions

13.2.1 Temporal Predictions

To plan appropriately requires predictions of where people will be at different points in time. More formally, we need predictions of expected future occupancy of each location during the time windows surrounding fixed intervals: $\tau, 2\tau, ..., T\tau$. We denote these quantities as $D_{s,t}$. In theory, time can be added to the state space of a Markov decision process and explicitly modeled. In practice, however, this expansion of the state space significantly increases the time complexity of inference, making real-time applications based on the time-based model impractical. We instead consider an alternative approach that is much more tractable.

We assume that a person's movement will "consume" some cost over a time window t according to the normal distribution $N(tC_0, \sigma_0^2 + t\sigma_t^2)$, where C_0, σ_0^2, and σ_t^2 are learned parameters. Certainly $\sum_t D_{s,t} = D_s$, so we simply divide the expected visitation counts among the time intervals according to this probability distribution. We use the cost of the optimal path to each state, $Q^*(s)$, to estimate the cost incurred in reaching it. The resulting time-dependent occupancy counts are then:

$$D_{s,it} \propto D_s e^{\frac{-(C_0 t - Q^*(s))^2}{2(\sigma_0^2 + t\sigma_1^2)}}. \tag{13.1}$$

These values are computed using a single execution of Dijkstra's algorithm (Dijkstra, 1959) in $O(|S| \log |S|)$ time to compute $Q^*(.)$ and then $O(|S|T)$ time for additional calculation.

Ideally, to account for predictive models of pedestrian behavior, we should increase the dimensionality of the planning problem by augmenting the state of the planner to account for time-varying costs. Unfortunately, the computational complexity of combinatorial planning is exponential in the dimension of the planning space, and the added computational burden of this solution will be prohibitive for many real-time applications.

We therefore propose a novel technique for integrating our time-varying predictions into the robot's planner. Algorithm 13.1 details this procedure; it essentially iteratively shapes a time-independent navigational cost function to remove known points of hindrance. At each iteration, we run the time-independent planner under the current cost map and simulate forward the resulting plan in order to predict points at which the robot will likely interfere with the pedestrian. By then adding cost to those regions of the map we can ensure that subsequent plans will be discouraged from interfering at those locations. We can further improve the computational gain of this technique by using efficient replanners such as D* and its variants (Ferguson & Stentz, 2005) in the inner loop. While this technique, as it reasons only about stationary costs, cannot guarantee the optimal plan given the time-varying costs, we demonstrate that it produces good robot behavior in practice that efficiently accounts for the predicted motion of the pedestrian.

By re-running this iterative replanner every 0.25 seconds using updated predictions of pedestrian motion, we can achieve intelligent adaptive robot behavior that anticipates where a pedestrian is heading and maneuvers well in advance to implement efficient avoidance. In practice, we use the final cost-to-go values of the iteratively constructed cost map to implement a policy that chooses a good action from a predefined collection of actions. When a plan with sufficiently low probability of pedestrian hindrance cannot be found, the robot's speed is varied. Additionally, when the robot is too close to a pedestrian, all actions that take the robot within a small radius of the human are removed to avoid potential collisions. Section 13.3.6 presents quantitative experiments demonstrating the properties of this policy.

13.3 Experimental Evaluation

We now present experiments demonstrating the capabilities of our prediction model and its usefulness for planning hindrance-sensitive robot trajectories.

13.3.1 Data Collection

We collected over one month's worth of data in a lab environment. The environment has three major areas (Figure 13.1): a kitchen area with a sink, refrigerator, microwave, and coffee maker; a secretary desk; and a lounge area. We installed four laser range finders in fixed locations around the lab, as shown in Figure 13.1, and ran a pedestrian tracking algorithm (MacLachlan, 2005). Trajectories were segmented based on significant stopping time in any location.

From the collected data, we use a subset of 166 trajectories through our experimental environment to evaluate our approach. This dataset is shown in Figure 13.3 after post-processing and

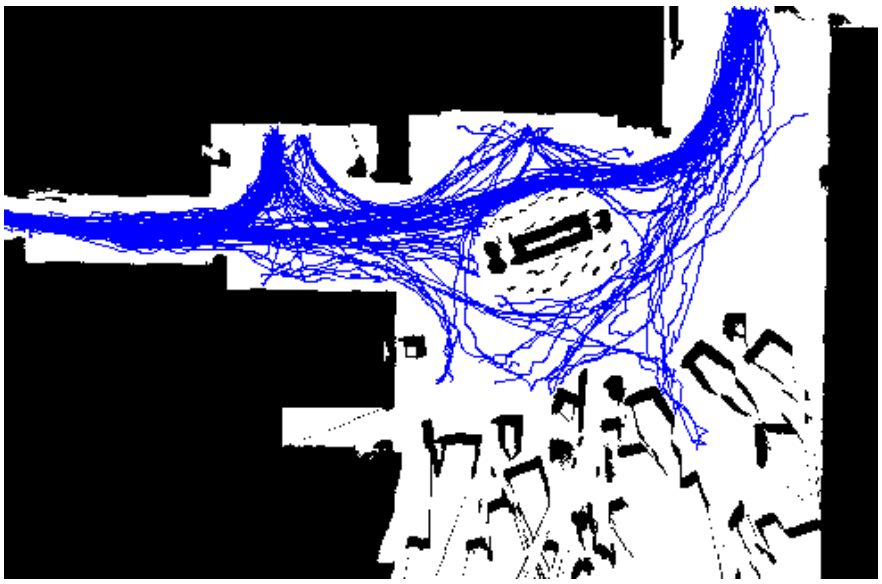

Figure 13.3: Collected trajectory dataset.

being fit to a 490 by 321 cell grid (each cell represented as a single pixel). We employ 50% of this data as a training set for estimating the parameters of our model and use the remainder for evaluative purposes.

13.3.2 Learning Feature-Based Cost Functions

We learn a 6-parameter cost function over simple features of the environment, which we argue is easily transferable to other environments. The first feature is a constant feature for every grid cell in the environment. The remaining functions are an indicator function for whether an obstacle exists in a particular grid cell, and four "blurs" of obstacle occupancies, which are shown in Figure 13.4.

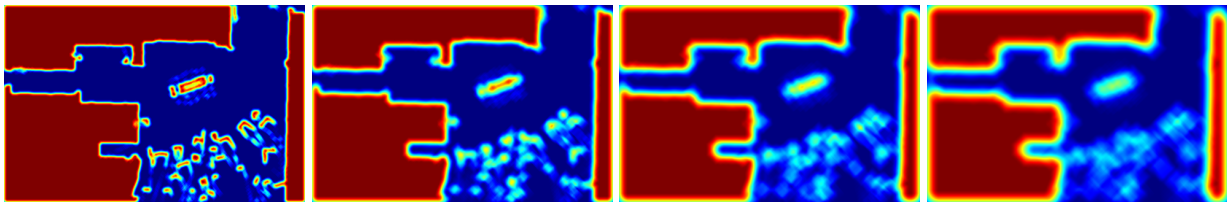

Figure 13.4: Four obstacle-blur features for our cost function. Feature values range from low weight (dark blue) to high weight (dark red).

We then learn the weights for these features that best explain the demonstrated data. The resulting cost function for the environment is shown in Figure 13.5. Obstacles in the cost function have

very high cost, and free space has a low cost that increases near obstacles. Essentially, the predictive model learns that travel through obstacles is extremely expensive and that travel near obstacles is costly, but not as expensive, relative to travel through open space. Despite the simplicity of the learned model, it generalizes well and provides useful predictions in practice.

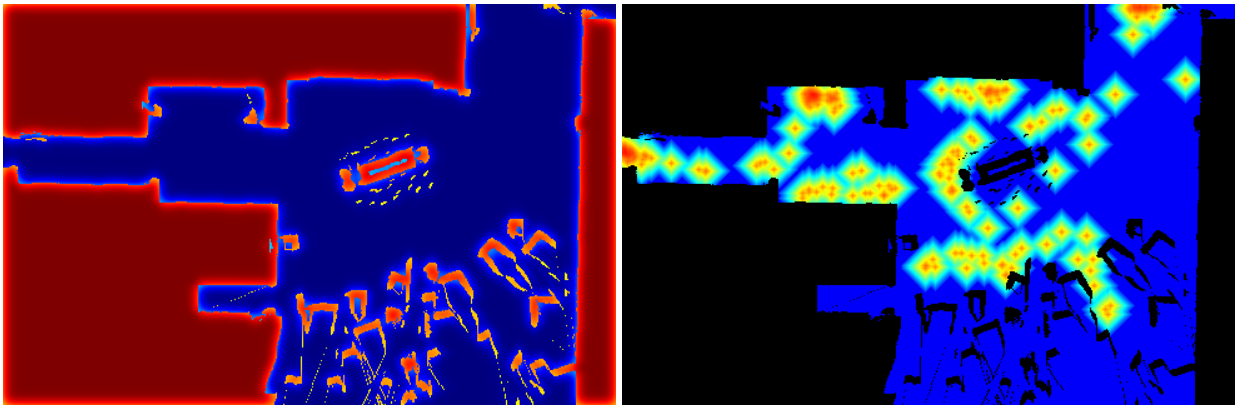

Figure 13.5: Left: The learned cost function in the environment. Right: The prior distribution over destinations learned from the training set.

The prior distribution over destinations is obtained from the set of endpoints in the training set, and the temporal Gaussian parameters are also learned using the training set.

13.3.3 Stochastic Modeling Experiment

We first consider two examples from our dataset (Figure 13.6) that demonstrate the need for uncertainty-based modeling.

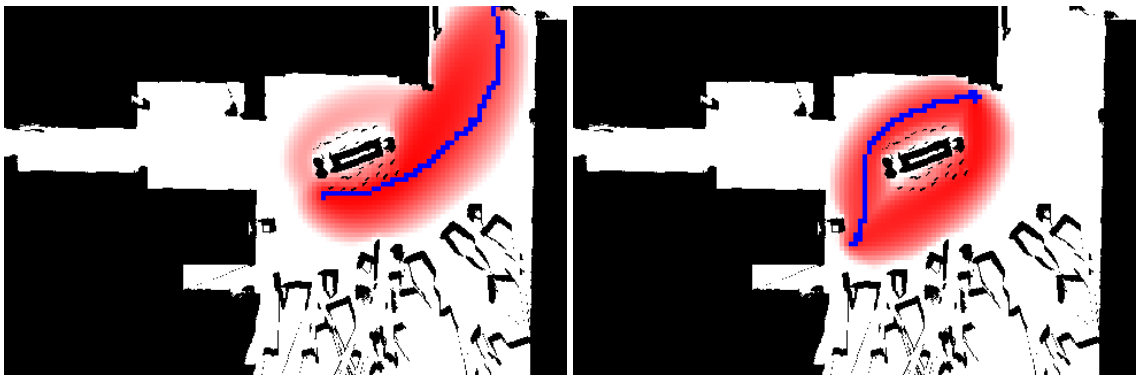

Figure 13.6: Two trajectory examples (blue) and log occupancy predictions (red).

Both trajectories travel around the table in the center of the environment. However, in the first example (left), the person takes the lower pathway around the table, and in the second example

(right), the person takes the upper pathway despite that the lower pathway around the table has a lower cost in the learned cost function. In both cases, the path taken is not the shortest path through the open space that one would obtain using an optimal planner. Our uncertainty-based planning model handles these two examples appropriately, while a planner would choose one pathway or the other around the table and, even after smoothing the resulting path into a probability distribution, tend to get a large fraction of its predictions wrong when the person takes the "other" approximately equally desirable pathway.

13.3.4 Dynamic Feature Adaptation Experiment

In many environments, the relevant features that influence movement change frequently: furniture is moved in indoor environments; the locations of parked vehicles are dynamic in urban environments; and weather conditions influence natural environments with muddy, icy, or dry conditions. We demonstrate qualitatively that our model of motion is robust to these feature changes.

The left frames of Figure 13.7 show the environment and the path prediction of a person moving around the table at two different points in time. At the second point of time (bottom left), the probability of the trajectory leading to the kitchen area or the left hallway is extremely small. In the right frames of Figure 13.7, an obstacle has been introduced that blocks the direct pathway through the kitchen area. In this case, the trajectory around the table (bottom right) still has a very high probability of leading to either the kitchen area or the left hallway. As this example shows, our approach is robust to changes in the environment such as this one.

13.3.5 Comparative Evaluation

We now compare our model's ability to predict the future path of a person with a previous approach for modeling goal-directed trajectories – the variable-length Markov model (VLMM) (Galata et al., 2001). The VLMM (introduced in Section 3.1.1) estimates the probability of a person's next cell transition conditioned on the person's history of cells visited in the past. It is variable length because it employs a long history when relevant training data is abundant, and a short history otherwise.

The results of our experimental evaluation are shown in Figure 13.8. We first note that for the training set (denoted *train*), that the trajectory log probability of the VLMM is significantly better than the plan-based model. However, for the test set, which is the metric we actually care about, the performance of the VLMM degrades significantly, while the degradation in the plan-based model is much less extreme. We conclude from this experiment that the VLMM (and similar directed graphical model approaches) are generally much more difficult to train to generalize well because their number of parameters is significantly larger than the number of parameters of the cost function employed in our approach.

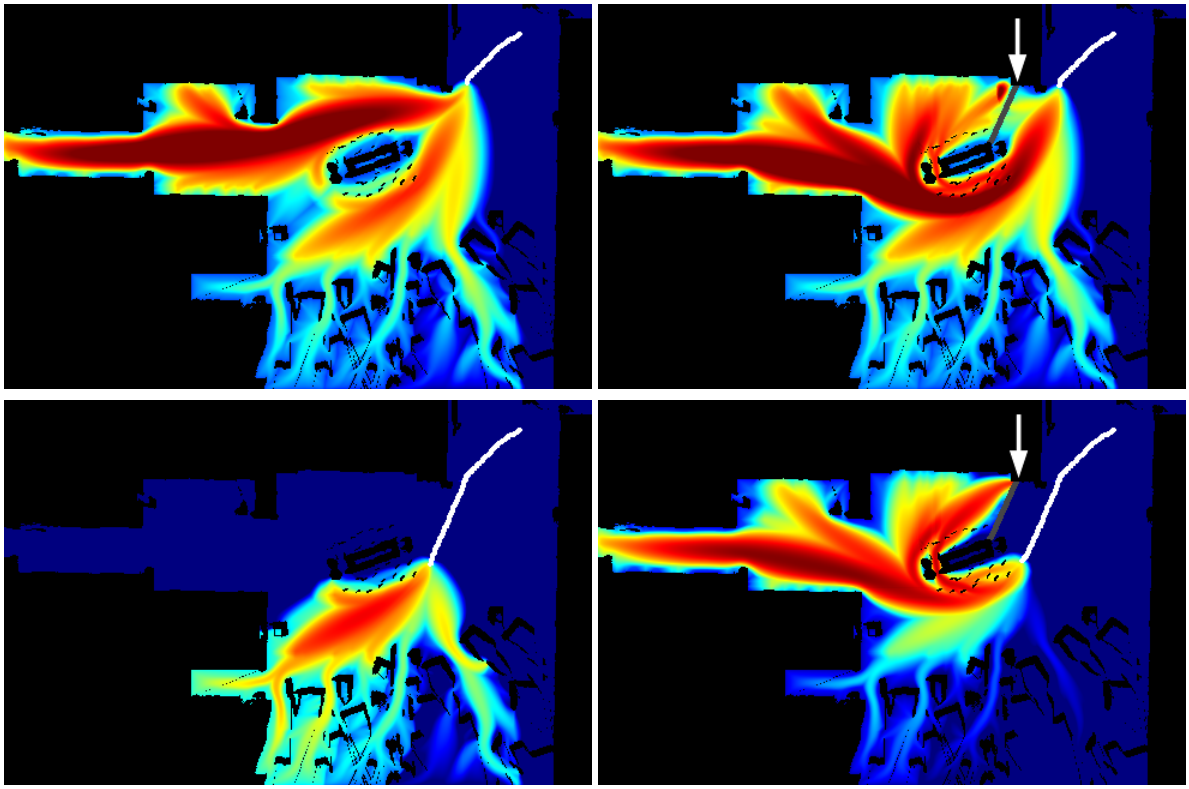

Figure 13.7: Our experimental environment with (right column) and without (left column) an added obstacle (gray, indicated by an arrow) between the kitchen and center table (with white arrow for emphasis). Predictions of future visitation expectations given a person's trajectory (white line) in both settings for two different trajectories. Frequencies range from red (high log expectation) to dark blue (low log expectation).

13.3.6 Integrated Planning Evaluation

We now simulate robot-human hindrance problems to demonstrate the benefit of our trajectory forecasting approach. We generate 200 hindrance-sensitive planning problems (corresponding to 22 different person trajectories in Figure 13.3) by selecting problems where naïve planning (disregarding the pedestrian) causes hindrances. We ignore the causal influence of the robot's action on the person's trajectory, and measure the trade-off between robot efficiency and interference with the person's trajectory. Specifically, the *average hindrance count* measures the average number of times the policy removed actions due to proximity to a human, and the *average execution time* measures the number of time steps needed to reach the goal. The trade-off is controlled by varying the degree of the visitation frequency threshold used in Algorithm 13.1. The robot trajectory planner is provided with person location forecasts at 4Hz.

The trade-off curves of planning using our plan-based forecasts and using a straight-forward particle-based forecasting model on this set of problems are shown in Figure 13.9. For both pre-

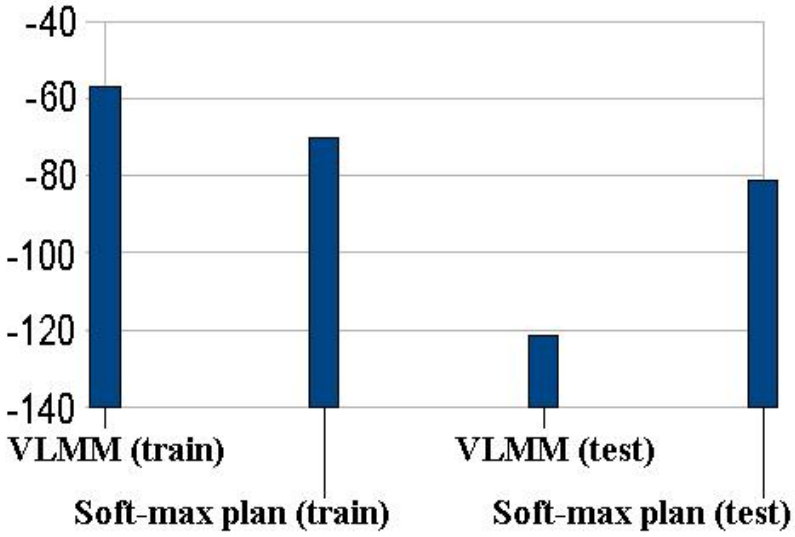

Figure 13.8: Log probability of datasets under the VLMM and our approach.

dictive models, while the execution time varies over a small range, the average hindrance count decreases by a factor of two. Additionally, as the figure shows, the plan-based forecasting is superior to the particle-based approach for almost any level of the trade-off.

13.4 Discussion

In this chapter, we have demonstrated the effective use of the maximum causal entropy approach for improving the navigation decisions of a robot. We employ a sophisticated predictions of the future positions of pedestrians within the robot's environment. The path planning employed by our approach does not employ inverse optimal control to learn the cost for robot navigation; it uses a fixed cost function to deter movements leading to hindrances.

Recently a complementary maximum entropy inverse optimal control has been employed to also improve robot navigation among people. This technical approach was extended to address settings with partial information. Whereas we use the inverse optimal control approach to model pedestrian behavior and employ the predictions in a cost function, Henry et al. (2010) pose the problem of path planning for a robot as a supervised problem and extend our maximum entropy approach to deal with partial observability of a probabilistic model of crowds of pedestrian movements. Combining these two approaches, or learning to adjust the robot's planner from interaction with pedestrians, is an important future direction for this line of research.

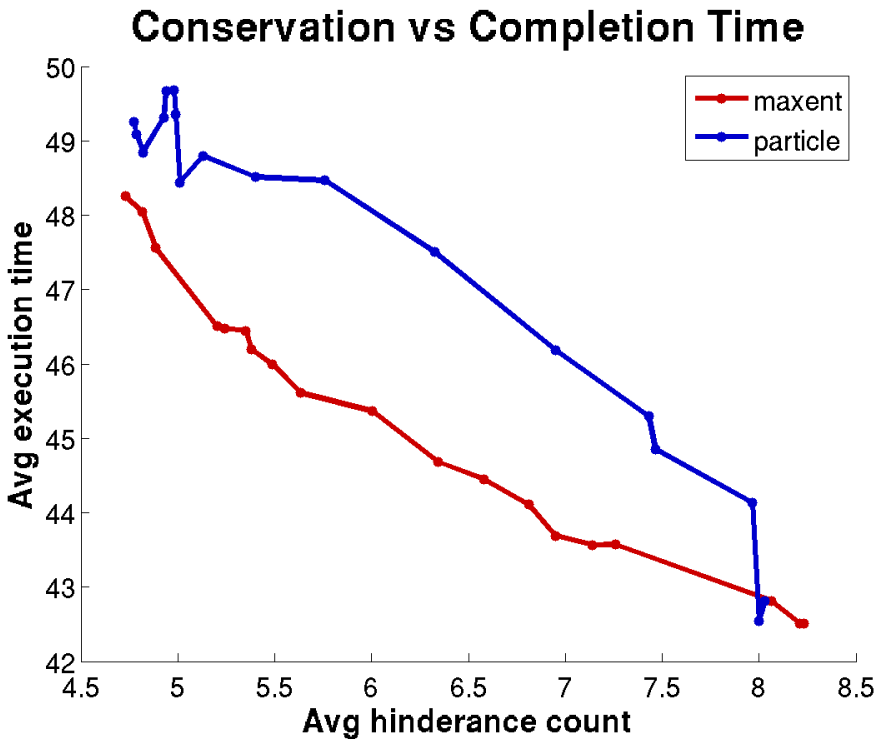

Figure 13.9: The trade-off in efficiency versus pedestrian hindrance for varying degrees of hindrance penalization for planning under both planning-based predictions and particle-based predictions.

Chapter 14

Other Applications

"Life is a sum of all your choices"
— Albert Camus (Philosopher, 1913–1960)

We now present three additional applications with varying complexity of interaction: the first is interactions of multiple players within a correlated equilibrium, the second is interaction with a partially observable system, and the third is inverse linear quadratic regulation.

14.1 Maximum Causal Entropy Correlated Equilibria for Markov Games

Using the theory developed in Chapter 8 for rational behavior in multi-player games, we present experiments for the robust prediction of rational behavior in stochastic games.

14.1.1 Experimental Setup

Following Zinkevich et al. (2006), we generate random stochastic Markov games according to the following procedure. For each of $|S|$ states in the Markov game, each of N players has $|A|$ actions to choose from (for a total of $|A|^N$ joint actions). The state transition dynamics, $\{\mathrm{P}(s_{t+1}|s_t, a_t^{1:N})\}$, depend on the combination of players' actions (and state) and are drawn uniformly from the simplex of probabilities. The utility obtained by each player in each state, $\text{Utility}_i(s)$, is drawn uniformly from $\{0, 0.1, 0.2, ..., 0.9\}$. For our experiments, we allow either the number of states to vary with 2 actions and 3 players; or the number of players to vary for random Markov games with 2 states and 2 actions. In both cases, we employ a discount factor of $\gamma = 0.75$. We consider a fixed time horizon of $T = 10$ time steps for our experiments.

We generate time-varying strategy profiles for MCECE using Algorithm 9.13 and for the CE-Q variants using projected sub-gradient optimization. The CE-Q strategies we evaluate are a subset

of those described in Section 8.1.2. *i*CEQ maximizes the positive margin of player 1's utility over player 2's utility. We repeat this process for 100 random games for each choice of game parameters and investigate the properties of the resulting CE strategy profiles.

14.1.2 Evaluation

While the joint entropy of trajectories through a game tends to greatly increase with the number of states due to increased transition dynamic stochasticity, the uncertainty of the per-state action uncertainty does not, as shown in Figure 14.1.

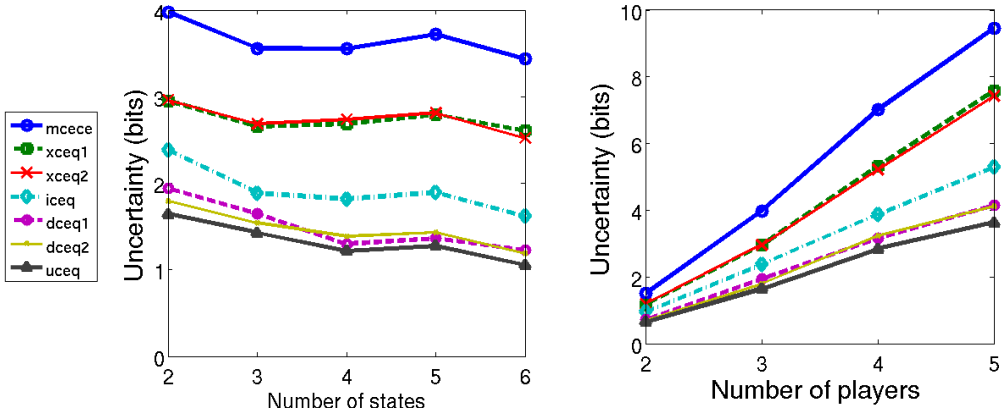

Figure 14.1: The ($\log_2$) causal entropy measure of the inherent difficulty of predicting the 10 time step action sequences that are generated by different correlated equilibria solution concepts' strategy profiles for varying numbers of states (left) with 2 actions and 3 players, and for varying numbers of players (right) with 2 actions and 2 states. It can be interpreted as the number of binary *bits* needed on average to represent the actions of a sampled trajectory under the strategy profile's distribution.

In fact, actions become slightly more predictable with more states, presumably due to the availability of more diverse action outcomes that better satisfy multiple players' objectives. By contrast, the uncertainty of action sequences increases logarithmically with the size of the action set, as one might expect. We note that in all cases, the MCECE strategy profile is the most uncertain (by design), and many of the previously investigated CE-Q solutions are the most deterministic.

The log probability of samples from one strategy profile under another strategy profile (*i.e.*, the log-loss or cross entropy),

$$-\sum_{a_{1:N}} \pi_A(a_{1:N}) \log \pi_B(a_{1:N}), \tag{14.1}$$

is a common, but harsh metric for evaluating the predictive capabilities of strategy profile B; it penalizes infinitely for actions with non-zero probability in A, but zero probability in B. Instead,

Table 14.1: The average cross-strategy-profile predictability for the single time step action distribution from the initial game state averaged over 100 random 3-player, 2-state, 2-action Markov games. Denoted are the percentage of strategies that are impossible to represent in the strategy profile (*i.e.*, having zero probability within the model). and the average number of bits needed to encode the combined action of those that are impossible to represent. We only count strategies with 0.1% probability or more under the predicted strategy profile for assessing representation possibility and never penalize more than 16.6 bits ($-\log_2 0.00001$) for any extremely small probability in the predicting strategy profile to avoid any approximation effects—both to the benefit of CE-Q strategy profiles.

	MCECE	*u*CE-Q	d_1CE-Q	d_2CE-Q	x_1CE-Q	x_2CE-Q	*i*CE-Q	Average
MCECE	———	1.951 0.0%	1.951 0.0%	1.967 0.0%	**2.010** 0.0%	**1.992** 0.0%	1.974 0.0%	**1.974** 0.0%
*u*CE-Q	3.377 22.3%	———	2.039 4.5%	**1.888** 5.7%	2.647 21.2%	3.072 20.8%	2.444 13.8%	2.578 14.7%
d_1CE-Q	3.442 18.9%	**1.866** 3.5%	———	2.511 5.6%	3.328 18.8%	2.321 17.6%	**1.798** 9.8%	2.544 12.4%
d_2CE-Q	3.462 17.0%	1.872 1.7%	2.536 3.3%	———	2.576 15.6%	3.489 17.2%	3.060 12.1%	2.833 11.2%
x_1CE-Q	2.897 0.2%	2.472 0.0%	2.798 0.0%	2.450 0.0%	———	2.375 0.0%	2.764 0.0%	2.626 0.0%
x_2CE-Q	**2.877** 0.5%	2.605 0.0%	2.251 0.0%	2.905 0.0%	2.373 0.2%	———	2.116 0.0%	2.521 0.1%
*i*CE-Q	3.378 5.3%	2.279 1.9%	**1.902** 0.0%	2.989 2.5%	3.116 5.1%	2.276 4.2%	———	2.657 3.2%

we extend the representational interpretation of the log probability and assess what fraction of samples in A are possible to represent in B and of the ones possible to represent, the average number of bits required to do so, $-\sum_{a_{1:N}:\pi_B(a_{1:N})>0} \pi_A(a_{1:N}) \log \pi_B(a_{1:N})$).

The results of this comparison across strategy profiles is shown in Table 14.1 (for 3 players) and Table 14.2 (for 4 players). We note that in some cases one C-EQ strategy profile may better predict another than the MCECE strategy profile when their objectives are closely aligned. For example, the x_2CE-Q tends to predict iCE-Q fairly accurately since both are punishing player 2 to some degree. However, overall the MCECE solution profile provides a much more robust prediction of other strategy profiles on average and full support of all possible action combinations.

14.2 Inverse Diagnostics

Many important problems can be framed as interaction with a partially observable stochastic system. In medical diagnosis, for example, tests are conducted, the results of which may lead to additional tests to narrow down probable conditions or diseases and to prescribe treatments, which

Table 14.2: The average cross-strategy-profile single action predictability for 4 players and otherwise the identical experimental setting as Table 14.1.

	MCECE	uCE-Q	d_1CE-Q	d_2CE-Q	x_1CE-Q	x_2CE-Q	iCE-Q	Average
MCECE	———	3.451 0.0%	**3.468** 0.0%	**3.476** 0.0%	**3.518** 0.0%	**3.509** 0.0%	3.475 0.0%	**3.483** 0.0%
uCE-Q	7.308 25.8%	———	3.955 9.4%	3.652 8.5%	6.495 21.8%	6.814 22.0%	5.591 17.9%	5.636 17.6%
d_1CE-Q	6.914 22.5%	3.831 5.8%	———	4.746 10.2%	7.566 21.2%	5.536 17.6%	3.895 11.2%	5.415 14.8%
d_2CE-Q	7.109 23.1%	**3.408** 6.4%	4.767 10.7%	———	5.643 18.3%	7.651 21.5%	6.976 19.7%	5.926 16.6%
x_1CE-Q	**4.603** 0.0%	4.300 0.0%	5.000 0.0%	3.849 0.0%	———	3.918 0.0%	4.372 0.0%	4.340 0.0%
x_2CE-Q	4.633 0.1%	4.401 0.0%	3.917 0.0%	4.986 0.0%	3.944 0.0%	———	**3.171** 0.0%	4.175 0.0%
iCE-Q	6.380 11.5%	5.070 5.0%	3.884 2.8%	6.501 8.0%	5.991 8.7%	4.311 5.5%	———	5.356 6.9%

are adjusted based on patient response. Motivated by the objective of learning good diagnosis policies from experts, we investigate the Inverse Diagnostics problem of modeling interaction with partially observed systems.

14.2.1 MaxCausalEnt ID Formulation

In this setting, the partially observed set of variables (related by a Bayesian Network in this application) serves as side information. Inference over the latent variables from this set is required to infer decision probabilities. Additionally, decisions can influence the variables, causally changing their values, and the implications of these interventions must also be assessed. Vectors of features are associated with observing or manipulating each variable, and we employ our MaxCausalEnt ID model with these features as value nodes as shown in Figure 14.2[1].

14.2.2 Fault Diagnosis Experiments

We apply our inverse diagnostics approach to the vehicle fault detection Bayesian Network (Heckerman et al., 1994) shown in Figure 14.3. Apart from the relationship between *Battery Age* and *Battery* (exponentially increasing probability of failure with battery age), the remaining condi-

[1] An objective function over the Bayesian Network variables can also be incorporated into a value node at each time-step and/or at the end of the sequence, as shown.

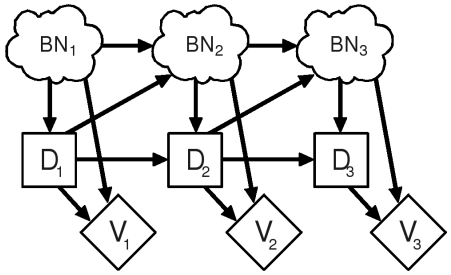

Figure 14.2: The MaxCausalEnt ID representation of the diagnostic problem.

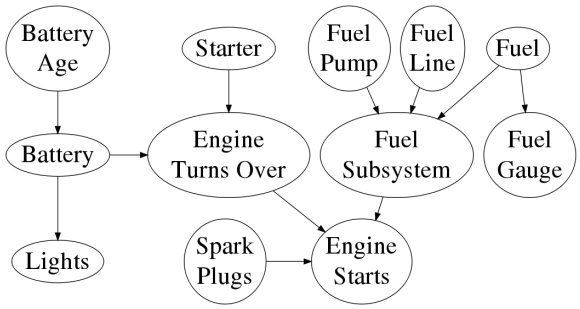

Figure 14.3: The vehicle fault detection Bayesian Network.

tional probability distributions are deterministic-or's (*i.e.*, failure in any parent causes a failure in the child). The remaining (conditional) probabilities is shown in Table 14.3.

Table 14.3: Probability distributions governing random variables in the vehicle fault diagnosis Bayesian network of the inverse diagnostics experiment.

$$
\begin{aligned}
P(\text{battery age} = x) &= \frac{1}{5} \text{ for } x = 0, 1, 2, 3, 4 \\
P(\text{starter failure}) &= 0.005 \\
P(\text{fuel pump failure}) &= 0.002 \\
P(\text{fuel failure}) &= 0.006 \\
P(\text{spark plug failure}) &= 0.003 \\
P(\text{battery failure}|\text{battery age}) &= 1 - e^{-.01 * \text{battery age}}
\end{aligned}
$$

Each variable in the network can be tested, revealing whether it is operational (or the battery's age), and the *Battery, Fuel, Fuel Line, Fuel Pump, Spark Plugs*, and *Starter* can each be replaced

(making it and potentially its descendants operational). Replacements and tests are both characterized *action features*: a cost to the vehicle owner, a profit for the mechanic, and a time requirement. Ideally a sequence of tests and replacements would minimize the expected cost to the vehicle owner, but an over-booked mechanic might instead choose to minimize the total repair time so that other vehicles can be serviced, and a less ethical mechanic might seek to optimize personal profit.

Table 14.4: Replacement and observation features for variables of the vehicle fault diagnosis Bayesian network. The first feature corresponds to an approximate cost to the vehicle owner. The second feature corresponds to an approximate profit to the mechanic. The final feature corresponds to a time requirement.

Variable	Replace Features	Observation Features
Battery Age		[5.0, 5.0, 0.0]
Starter	[120.0, 100.0, 40.0]	[60.0, 30.0, 20.0]
Fuel Pump	[150.0, 120.0, 30.0]	[70.0, 40.0, 20.0]
Fuel Line	[50.0, 30.0, 15.0]	[40.0, 20.0, 15.0]
Fuel	[30.0, 25.0, 10.0]	[15.0, 10.0, 5.0]
Batter	[140.0, 120.0, 20.0]	[50.0, 30.0, 20.0]
Engine Turns Over		[5.0, 5.0, 3.0]
Fuel Subsystem		[10.0, 10.0, 5.0]
Fuel Gauge		[3.0, 3.0, 2.0]
Lights		[2.0, 2.0, 2.0]
Spark Plugs	[90.0, 60.0, 40.0]	[50.0, 30.0, 20.0]
Engine Starts		[5.0, 5.0, 3.0]

To generate a dataset of observations and replacements, a stochastic policy is obtained by adding Gaussian noise, $\epsilon_{s,a}$, to each action's future expected value, $Q^*(s, a)$, under the optimal policy for a fixed set of weights and selecting the highest noisy-valued action, $Q^*(s, a) + \epsilon_{s,a}$, to execute at each time-step. Different vehicle failure samples are generated from the Bayesian Network conditioned on the vehicle's engine failing to start, and the stochastic policy is sampled until the vehicle is operational.

We evaluate the prediction error rate and log-loss of our model in Figure 14.4. We compare against a Markov Model that ignores the underlying mechanisms for decision making and simply predicts behavior in proportion to the frequency it has previously been observed (with small pseudo-count priors). Our approach consistently outperforms the Markov Model even with an order of magnitude less training data. The classification error rate quickly reaches the limit implied by the inherent stochasticity of the data generation process.

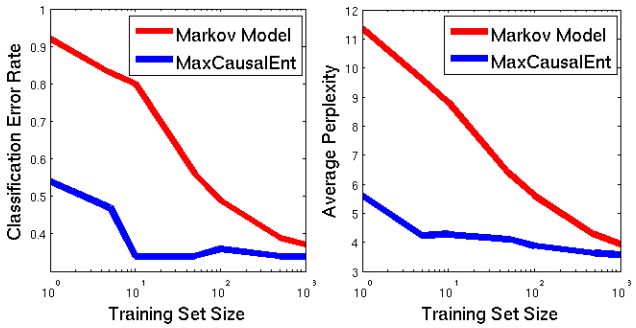

Figure 14.4: Error rate and log-loss of the MaxCausalEnt ID model and Markov Model for diagnosis action prediction as training set size (log-scale) increases.

14.3 Helicopter Control

We demonstrate the MaxCausalEnt approach to inverse stochastic optimal control on the problem of building a controller for a learned helicopter model (Abbeel et al., 2007) with linearized stochastic dynamics. Most existing approaches to inverse optimal control with stochastic dynamics (Ratliff et al., 2006; Abbeel & Ng, 2004) have both practical and theoretical difficulties in the presence of imperfect demonstrated behavior, leading to unstable controllers due to large changes in cost weights (Abbeel et al., 2007) or poor predictive accuracy (Ratliff et al., 2006).

14.3.1 Experimental Setup

To test the robustness of our approach, we generated five 100 timestep sub-optimal training trajectories by noisily sampling actions from an optimal LQR controlled designed for hovering using the linearized stochastic simulator of Abbeel et al. (2007). It simulates a XCell Tempest helicopter using four continuous control variables: elevator, aileron, rudder, main rotor; and a 21 dimensional state space: position, orientation, velocity, angular rate, previous controls, and change in previous controls.

14.3.2 Evaluation

We contrast performance between maximum margin planning (Ratliff et al., 2006) (labeled InvOpt in Figure 14.5) and MaxCausalEnt trained using demonstrated trajectories. Performance was evaluated by generating trajectories from the optimal controller of each model and measuring their cost under the *true* cost function used to generate the original sub-optimal demonstration trajectories. The InvOpt model performs poorly because there is no optimal trajectory *for any cost function* that matches demonstrated features.

On the other hand, the function learned by MaxCausalEnt IOC not only induces the same

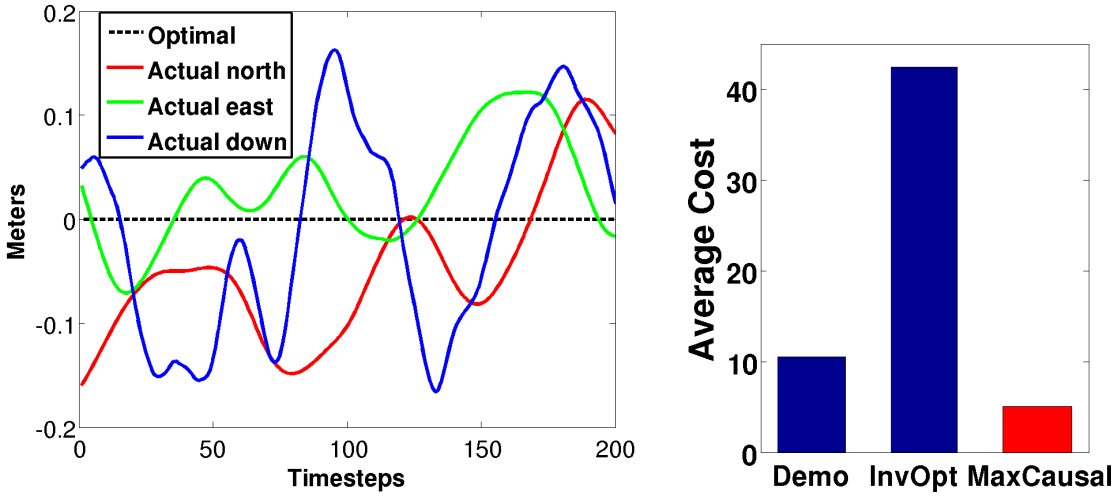

Figure 14.5: Left: An example sub-optimal helicopter trajectory attempting to hover around the origin point. Right: The average cost under the original cost function of: (1) demonstrated trajectories; (2) the optimal controller using the inverse optimal control model; and (3) the optimal controller using the maximum causal entropy model.

feature counts– and hence equal cost on the unknown cost function– under the learned probabilistic policy, but because of the quadratic cost function its learned controller's optimal policy is always *at least as good* as the demonstrated behavior on the original, unknown cost function. Figure 14.5 demonstrates this; the resulting learned optimal policy outperforms the demonstrated behavior on the original, unknown cost function. In this sense, MaxCausalEnt provides a rigorous approach to learning a cost function for such stochastic optimal control problems: it is both predictive and can guarantee good performance of the learned controller.

14.4 Discussion

In this chapter, we demonstrated three applications of the maximum causal entropy approach that extend beyond discrete-valued inverse optimal control. Specifically, we investigated the setting of multiple rational agents, the setting with additional latent information, and the setting of continuous control. These applications help to illustrate the versatility of the maximum causal entropy principle.

Part V

Conclusions

Chapter 15

Open Problems

"Science never solves a problem without creating ten more."
— George Bernard Shaw (Playwright, 1856–1950).

Throughout this thesis, we have made numerous assumptions about the structural relationships among variables and the conditional probability distributions from which side information is drawn. Specifically, we have assumed that the structure and the probability distributions are known—both to the agent acting to generate behavior sequences, and to the observer trying to model and predict the agent's future behavior—, that certain restrictions exist on the influence of decisions and side information on one another, and that deviation regret constrains the distribution of actions. Many possible extensions to the presented work are based on relaxing these assumptions in various ways.

15.1 Structure Learning and Perception

Throughout this thesis, we have assumed that the dynamics governing side information and which side information influences which decisions are both known. However, in many problem settings, only observed data from a sequential interaction with a stochastic process is available, and learning both the structural dependencies between variables and probability distributions relating observed data is a necessary task. Importantly, in the maximum causal entropy model a sequence of variables causally conditioned on another sequence of variables is distributed very differently than a sequence statically conditioned on another sequence.

Learning both the structure and distributions relating side information and decision variables is an important direction for future research. In the context of decision making, this is a problem of modeling the perception process of the decision maker. A number of important questions follow:

- What classes of structures are indistinguishable from one another given only observational data?

- How efficiently in terms of sample sizes can distinguishable classes of structures be reliably learned from observational data drawn from a distribution and structure that is in the class of the maximum causal entropy framework?

- How computationally efficiently can maximum causal entropy structure learning be performed?

- What approximation techniques can be employed to make maximum causal entropy structure learning more efficient?

- How can theories of perception be incorporated into the structure learning process?

Assuming that the agent knows the structure and distribution of side information variables (but the learner does not) provides a useful starting point. Relaxing that assumption of agent knowledge of side information probabilities (and even structure) leads to interesting areas of new research where the maximum causal entropy approach is likely to serve a foundational role.

15.2 Predictive Strategy Profiles

For multi-player settings, existing equilibria solution concepts are largely defined for *prescriptive* settings where the actions of an autonomous player (or players) needs to be obtained and employed within a game setting. For *predictive* settings, most equilibria solution concepts are underspecified—typically they allow multiple solutions and no probabilistic bias over the set of valid equilibria to employ as a probability distribution—making them inapplicable without additional assumptions. Additionally, in numerous experiments (McKelvey & Palfrey, 1992), the actions of human game players often do not follow any rational equilibria solution that game theory prescribes.

Predictive models of players' strategies relax the perfect rationality assumptions of game theory to allow for bounded rationality, error-prone action selection, and ulterior motives beyond the game payoff.

Definition 15.1. *In a **quantal response equilibrium** (McKelvey & Palfrey, 1995, 1998), players actions are noisily distributed based on the expected utility of the action, typically according to a logit response function. For normal form games, player i's action choice is distributed according to:*

$$\mathrm{P}(a_i) = \frac{e^{\lambda \, \mathrm{E}_{\mathrm{P}(\boldsymbol{a})}[\mathit{Utility}_i(\boldsymbol{a}) | a_i]}}{\sum_{a'_i} e^{\lambda \, \mathrm{E}_{\mathrm{P}(\boldsymbol{a})}\left[\mathit{Utility}_i(\boldsymbol{a}) | a'_i\right]}}. \tag{15.1}$$

In the extension to Markov games, the future expected utility is computed based on the assumption that the quantal response equilibrium is also played in the future:

$$\mathrm{P}(a_i^t) = \frac{e^{\lambda \, \mathrm{E}_{\mathrm{P}(\boldsymbol{a}^t)}\left[\mathit{EUtility}_i^t(\boldsymbol{a}^t) | a_i^t\right]}}{\sum_{a_i^{t\prime}} e^{\lambda \, \mathrm{E}_{\mathrm{P}(\boldsymbol{a}^t)}\left[\mathit{EUtility}_i(\boldsymbol{a}^t) | a_i^{t\prime}\right]}}. \tag{15.2}$$

A desirable quality of the quantal response equilibrium (Definition 15.1) is that the parameter λ controls the rationality of the distribution. Namely, as λ goes to infinity, the distribution converges to a Nash equilibrium.

An alternative model of strategy profiles more explicitly models the bounded rationality of human players.

Definition 15.2. *In a **level-k** model of behavior (Costa-Gomes et al., 2001), a level-k agent plays the optimal response strategy profile against a level-$(k-1)$ player. A level-0 player chooses actions uniformly at random.*

Recent quantitative evaluations (Wright & Leyton-Brown, 2010) have shown that the combination of these two approaches (Stahl & Wilson, 1995) with the model learned and evaluated through a cross-validation procedure provide significantly more accurate predictions than either model alone.

Establishing some form of maximum entropy duality of these existing approaches and generalizing to the setting where features rather than payoffs are known are both important open problems. The formulation, however, must be nuanced; simply enforcing the feature-matching constraints of inverse optimal control can often lead to predicted behaviors that are highly random and only weakly adversarial when given behavior that is highly purposeful and very competitive.

15.3 Closing the Prediction-Intervention-Feedback Loop

Creating a closed-loop assistive system that predicts the intentions of human behaviors, executes interventions on behalf of its users, and assesses feedback from the user to self-improve the entire system is the ultimate application goal for the research of this thesis. The predictive techniques, algorithms, and applications described in earlier chapters are one important component of realizing such a system. While we believe that accurate behavior predictions are crucial for selecting appropriate interventions, learning the context-dependent appropriateness of the possible interventions in terms of those behavior predictions is another important learning problem.

The assistive navigation system application provides a natural testbed for creating closed-loop interventional feedback system. Many additional interesting settings exist for a such a closed-loop system. If an end-task utility is defined according to a known measure, reinforcement learning techniques can be employed to find intervention strategy that best improve end-task utility. Importantly, the adaptation of the user to the intervention strategy is an important aspect to consider.

When task-level feedback is not as explicit, user adaptation alone can be leveraged to understand which intervention strategies are desirable to the user.

Chapter 16

Conclusions and Discussion

"Essentially, all models are wrong, but some are useful."
— George Box (Statistician, 1919–).

We conclude by summarizing and discussing the theoretical and empirical arguments we have presented in this thesis in support of its central claim:

> **The principle of maximum causal entropy creates probabilistic models of decision making that are purposeful, adaptive, and/or rational, providing more accurate prediction of human behavior.**

16.1 Matching Purposeful Characteristics

The premise of our approach has been that many human behaviors are purposeful and that there are measurable characteristics that differentiate purposeful behavior from non-purposeful behavior. We see this in the vehicle navigation modeling application in the form of destination and road characteristics that define the desirability (or lack thereof) of different road segments of the road network. In the pedestrian trajectory prediction work, we saw this in the avoidance of obstacles and areas near obstacles. Statistical models capable of matching those purposeful characteristics are desired, but they must be capable of generalizing beyond the behavior previously seen.

16.2 Information-Theoretic Formulation

The principle of maximum causal entropy provides the important predictive guarantees of minimax log-loss for prediction in the sequential setting. This establishes it as a principled approach for modeling sequential data in a range of decision, control, and game settings where the future is not completely controllable. For this guarantee to be of practical importance, the constraints employed

within the maximum causal entropy framework must represent the essence of underlying factors influencing the distribution. In the context of behavior modeling, we have focused on constraints on the terminal or goal characteristics of behavior sequences and basis on functions describing states and actions that can be interpreted as utility or cost potential components.

16.3 Inference as Softened Optimal Control

To support the purposeful viewpoint, we have established a close algorithmic connection relating inference in our proposed maximum causal entropy framework to optimal behavior in decision theory, planning, and control. Inference in the maximum causal entropy framework can then be viewed as a softened version of the mechanisms employed for optimal decision making. We consider this connection to be significant in demonstrating that the models we have developed are inherently biased towards purposeful behavior. Much of the learning problem is then learning the utility function (*i.e.*, cost or reward) that guides behavior sequences in this relaxed analogy of the optimal decision making model.

16.4 Applications Lending Empirical Support

We have evaluated the maximum causal entropy approach on prediction tasks in personal navigation, pedestrian movement, assistive control, and multiplayer stochastic games to demonstrate its performance in comparison to existing approaches. We have shown improved performance over existing state-of-the-art techniques for:

- Predicting route selections compared with directed graphical models in the vehicle navigation domain.
- Predicting pedestrian movements compared with filtering-based approaches in the hindrance-sensitive robot planning work.
- Robustly predicting correlated equilibria strategy profiles in Markov games settings.

Additionally, we have shown simple experiments demonstrating the benefits of the approach on the inverse diagnostics setting where side information is latent and on the inverse linear quadratic regulation setting with continuous states and actions.

Appendix A

Proofs

A.1 Chapter 4 Proofs

Theorem 4.22. *Generally, the conditional and causal entropies are related as:*

$$H(\boldsymbol{Y}|\boldsymbol{X}) \leq H(\boldsymbol{Y}^T||\boldsymbol{X}^T) \leq H(\boldsymbol{Y}) \leq H(\boldsymbol{Y}, \boldsymbol{X}), \tag{A.1}$$

since additional conditioning can never increase entropy. In the special case that $\mathrm{P}(\boldsymbol{X}^T||\boldsymbol{Y}^{T-1})$ *is a deterministic function, then:*

$$H(\boldsymbol{Y}^T||\boldsymbol{X}^T) = H(\boldsymbol{Y}) = H(\boldsymbol{Y}, \boldsymbol{X}). \tag{A.2}$$

Proof. If $\mathrm{P}(\mathbf{X}^T||\mathbf{Y}^{T-1})$ is deterministic, $H(\mathbf{X}^T||\mathbf{Y}^{T-1}) = 0$. By definition, $H(\mathbf{Y}, \mathbf{X}) = H(\mathbf{Y}^T||\mathbf{X}^T) + H(\mathbf{X}^T||\mathbf{Y}^{T-1})$, and since $H(\mathbf{X}^T||\mathbf{Y}^T) = 0$, then $H(\mathbf{Y}, \mathbf{X}) = H(\mathbf{Y}^T||\mathbf{X}^T)$. □

A.2 Chapter 5 Proofs

Theorem 5.8. *The general maximum causal entropy optimization (Equation 5.6)—or, more correctly, minimizing its negation,*

$$\begin{aligned} &\underset{\{\mathrm{P}(Y_t|\boldsymbol{X}_{1:t}, \boldsymbol{Y}_{1:t-1})\}}{\operatorname{argmin}} \quad -H(\boldsymbol{Y}^T||\boldsymbol{X}^T) \\ &\textit{such that: } \mathrm{E}_{\mathrm{P}(\boldsymbol{X},\boldsymbol{Y})}[\mathcal{F}(\boldsymbol{X}, \boldsymbol{Y})] = \tilde{\mathrm{E}}_{\mathrm{P}(\boldsymbol{X},\boldsymbol{Y})}[\mathcal{F}(\boldsymbol{X}, \boldsymbol{Y})]; \\ &\forall_{\boldsymbol{X}_{1:t}, \boldsymbol{Y}_{1:t-1}} \sum_{Y_t} \mathrm{P}(Y_t|\boldsymbol{X}_{1:t}, \boldsymbol{Y}_{1:t-1}) = 1 \end{aligned} \tag{A.3}$$

—is a convex optimization. Further, when a feasible solution exists on the relative interior of the constraints, strong duality holds.

Proof. While we assumed the form of $\mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T)$ as a function of $\{\mathrm{P}(Y_t|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t-1})\}$ in Equation A.3, our optimization can be equivalently posed as a function of $\{\mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T)\}$ with linear constraints forcing normalization that prevents acausal influence of future side information, as shown:

$$\begin{aligned} &\underset{\{\mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T)\}}{\operatorname{argmin}} \; -H(\mathbf{Y}^T||\mathbf{X}^T) \\ \text{such that: } &\mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})}[f_k(\mathbf{X},\mathbf{Y})] = \mathrm{E}_{\tilde{\mathrm{P}}(\mathbf{X},\mathbf{Y})}[f_k(\mathbf{X},\mathbf{Y})] \;\text{ for } k = 1, ..., K; \\ &\forall_{\mathbf{X}} \sum_{\mathbf{Y}} \mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T) = 1; \\ \text{and: } &\forall_{\mathbf{X},\hat{\mathbf{X}}:\mathbf{X}_{1:\tau}=\hat{\mathbf{X}}_{1:\tau}} \sum_{\mathbf{Y}_{\tau+1:T}} \mathrm{P}(\mathbf{Y}||\mathbf{X}) = \sum_{\mathbf{Y}_{\tau+1:T}} \mathrm{P}(\mathbf{Y}||\hat{\mathbf{X}}) \end{aligned} \tag{A.4}$$

Starting with $x_i \log x_i$, which is a known convex function for $x_i > 0$ (since $\frac{\partial^2}{\partial x_i^2} x_i \log x_i = \frac{1}{x_i} \geq 0$), taking a positive linear combination of many of these terms preserves convexity, and yields: $\sum_i C_i x_i \log x_i$. $-H(\mathbf{Y}^T||\mathbf{X}^T)$ follows this form, and is thus convex. All the constraints of Equation A.4 are linear functions of $\mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T)$. □

Theorem 5.9. *Strong duality holds for the maximum causal entropy optimization.*

Proof. Strong duality is guaranteed if *Slater's condition* holds (Boyd & Vandenberghe, 2004). Slater's condition applied to this optimization is: There exists $\{\mathrm{P}(\mathbf{Y}^T||\mathbf{X}^T)\}$ that is *strictly feasible* on the probabilistic inequality constraints. This condition fails when:

1. Matching features is infeasible; or

2. The distribution is singular and does not have full support.

As long as the model and empirical data are drawn from the same distribution of $\mathrm{P}(\mathbf{X}^T||\mathbf{Y}^{T-1})$, condition (1) is satisfied. Condition (2) corresponds to an extreme point. It can be avoided by loosening the constraints by ϵ to guarantee full support. Since this extreme point corresponds to perfectly predictable behavior, a deterministic behavior distribution will be able to match the constraints without requiring slack. □

Theorem 5.10. *The maximum causal entropy distribution minimizes the worst case prediction log-loss,*

$$\inf_{\mathrm{P}(\boldsymbol{Y}||\boldsymbol{X})} \sup_{\tilde{\mathrm{P}}(\boldsymbol{Y}^T||\boldsymbol{X}^T)} -\sum_{\boldsymbol{Y},\boldsymbol{X}} \tilde{\mathrm{P}}(\boldsymbol{Y},\boldsymbol{X}) \log \mathrm{P}(\boldsymbol{Y}^T||\boldsymbol{X}^T),$$

given that $\tilde{\mathrm{P}}(\boldsymbol{Y}, \boldsymbol{X}) = \tilde{\mathrm{P}}(\boldsymbol{Y}^T||\boldsymbol{X}^T)\,\mathrm{P}(\boldsymbol{X}^T||\boldsymbol{Y}^{T-1})$ *and feature expectations* $\mathrm{E}_{\tilde{\mathrm{P}}(\boldsymbol{X},\boldsymbol{Y})}[\mathcal{F}(\boldsymbol{X}, \boldsymbol{Y})]$ *when* $\boldsymbol{X}$ *variables are sequentially revealed.*

Proof (sketch). The theorem is essentially a special case of Grünwald & Dawid's generalized maximum entropy–robust Bayes duality (2003). We provide the basic outline for the *maximin* and *minimax* interpretations of maximum causal entropy and refer the reader to Grünwald & Dawid's paper for additional details.

The causal entropy can be defined as: $H(\tilde{\mathrm{P}}(\mathbf{Y}||\mathbf{X})) = \inf_{\mathrm{P}(\mathbf{Y}||\mathbf{X})} \mathrm{E}_{\tilde{\mathrm{P}}(\mathbf{X},\mathbf{Y})}[-\log \mathrm{P}(\mathbf{Y}||\mathbf{X})]$. Choosing $\tilde{\mathrm{P}}(\mathbf{Y}||\mathbf{X})$ to maximize this is then: $\sup_{\tilde{\mathrm{P}}(\mathbf{Y}||\mathbf{X})} \inf_{\mathrm{P}(\mathbf{Y}||\mathbf{X})} \mathrm{E}_{\tilde{\mathrm{P}}(\mathbf{X},\mathbf{Y})}[-\log \mathrm{P}(\mathbf{Y}||\mathbf{X})]$. As a consequence of convex duality, this final expression is equivalent to the worst-case log-loss of the theorem (by swapping maximization and minimization ordering). □

A.3 Chapter 6 Proofs

Theorem 6.2. *The distribution satisfying the maximum causal entropy constrained optimization (Equation A.3) is recursively defined as:*

$$\mathrm{P}_\theta(Y_t|\boldsymbol{X}_{1:t}, \boldsymbol{Y}_{1:t-1}) = \frac{Z_{Y_t|\boldsymbol{X}_{1:t},\boldsymbol{Y}_{1:t-1},\theta}}{Z_{\boldsymbol{X}_{1:t},\boldsymbol{Y}_{1:t-1},\theta}} \tag{A.5}$$

$$\log Z_{\boldsymbol{X}_{1:t},\boldsymbol{Y}_{1:t-1},\theta} = \log \sum_{Y_t} Z_{Y_t|\boldsymbol{X}_{1:t},\boldsymbol{Y}_{1:t-1},\theta} = \operatorname*{softmax}_{Y_t} \Big(\sum_{X_{t+1}} \mathrm{P}(X_{t+1}|\boldsymbol{X}_{1:t}, \boldsymbol{Y}_{1:t}) \log Z_{\boldsymbol{X}_{1:t+1},\boldsymbol{Y}_{1:t},\theta} \Big)$$

$$Z_{Y_t|\boldsymbol{X}_{1:t},\boldsymbol{Y}_{1:t-1},\theta} = e^{\sum_{X_{t+1}} \mathrm{P}(X_{t+1}|\boldsymbol{X}_{1:t},\boldsymbol{Y}_{1:t}) \log Z_{\boldsymbol{X}_{1:t+1},\boldsymbol{Y}_{1:t},\theta}}$$

$$Z_{\boldsymbol{X}_{1:T},\boldsymbol{Y}_{1:T-1},\theta} = e^{\theta^\top \mathcal{F}(\boldsymbol{X},\boldsymbol{Y})},$$

where $\operatorname{softmax}_x f(x) \triangleq \log \sum_x e^{f(x)}$.

Proof. Differentiating the Lagrangian of the maximum causal entropy optimization (Equation A.3),

$$\Lambda(\mathrm{P}, \theta) = H(\mathbf{Y}^T||\mathbf{X}^T) + \sum_k \theta_k \Big(\mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})}[\mathcal{F}_k(\mathbf{X}, \mathbf{Y})] - \mathrm{E}_{\tilde{\mathrm{P}}(\mathbf{X},\mathbf{Y})}[\mathcal{F}_k(\mathbf{X}, \mathbf{Y})] \Big) + \sum_{t,\mathbf{X}^{1:t},\mathbf{Y}^{1:t-1}} C_{\mathbf{X}^{1:t},\mathbf{Y}^{1:t-1}} \left(\sum_{Y_t} \mathrm{P}(Y_t|\mathbf{X}^{1:t}, \mathbf{Y}^{1:t-1}) - 1 \right), \tag{A.6}$$

we have:

$$\nabla_{\{\mathrm{P}(Y_t|\mathbf{X}_{1:t},\mathbf{Y}_{1:t-1})\}} \Lambda(\mathrm{P}, \theta) = \Big\{ C_{\mathbf{X}_{1:t},\mathbf{Y}_{1:t-1}} - \mathrm{P}_\theta(\mathbf{X}_{1:t}, \mathbf{Y}_{1:t-1}) \tag{A.7}$$

$$\Big(\sum_{\tau=t}^{T} \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})}[\log \mathrm{P}_\theta(Y_\tau|\mathbf{X}_{1:\tau}, \mathbf{Y}_{1:\tau-1})|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t}] + \sum_k \theta_k \, \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})}[F_k(\mathbf{X}, \mathbf{Y})|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t}] \Big) \Big\}$$

Note that at this point we can pull the $\log \mathrm{P}_\theta(Y_t|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t-1})$ term back out of the expectation, equate the gradient to 0, and solve, yielding:

$$\mathrm{P}_\theta(Y_t|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t-1}) \propto e^{\theta^\top \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})}[\mathcal{F}(\mathbf{X},\mathbf{Y})|\mathbf{X}_{1:t},\mathbf{Y}_{1:t}] - \sum_{\tau=t+1} \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})}[\log \mathrm{P}_\theta(Y_\tau|\mathbf{X}_{1:\tau},\mathbf{Y}_{1:\tau-1})|\mathbf{X}_{1:t},\mathbf{Y}_{1:t}]}. \quad \text{(A.8)}$$

Using this recursive definition, we go further to prove the operational recurrence of the theorem. We will ignore the $C_{\mathbf{X}_{1:t},\mathbf{Y}_{1:t-1}}$ term for now, ultimately setting it to the remaining normalization term after factoring out the $\mathrm{P}_\theta(\mathbf{X}_{1:t}, \mathbf{Y}_{1:t-1})$ multiplier and substituting our recursive definitions.

$$\begin{aligned}
&- \sum_{\tau=t}^{T} \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})} \left[\log \mathrm{P}_\theta(Y_\tau|\mathbf{X}_{1:\tau}, \mathbf{Y}_{1:\tau-1}) \middle| \mathbf{X}_{1:t}, \mathbf{Y}_{1:t} \right] + \theta^\top \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})} \left[\mathcal{F}(\mathbf{X}, \mathbf{Y}) \middle| \mathbf{X}_{1:t}, \mathbf{Y}_{1:t} \right] \\
= &- \sum_{\tau=t}^{T-1} \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})} \left[\sum_{X_{\tau+1}} \mathrm{P}_\theta(X_{\tau+1}|\mathbf{X}_{1:\tau}, \mathbf{Y}_{1:\tau}) \log Z_\theta(\mathbf{X}_{1:\tau+1}, \mathbf{Y}_{1:\tau}) - \log Z_\theta(\mathbf{X}_{1:\tau}, \mathbf{Y}_{1:\tau-1}) \middle| \mathbf{X}_{1:t}, \mathbf{Y}_{1:t} \right] \\
&\quad - \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})}[\theta^\top \mathcal{F}(\mathbf{X}, \mathbf{Y}) - \log Z_\theta(\mathbf{X}_{1:T}, \mathbf{Y}_{1:T-1})|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t}] + \theta^\top \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})} \left[\mathcal{F}(\mathbf{X}, \mathbf{Y}) \middle| \mathbf{X}_{1:t}, \mathbf{Y}_{1:t} \right] \\
= &- \sum_{\tau=t}^{T} \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})} \left[\log Z_\theta(\mathbf{X}_{1:\tau+1}, \mathbf{Y}_{1:\tau}) - \log Z_\theta(\mathbf{X}_{1:\tau}, \mathbf{Y}_{1:\tau-1}) \middle| \mathbf{X}_{1:t}, \mathbf{Y}_{1:t} \right] \\
&\quad - \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})}[\log Z_\theta(\mathbf{X}_{1:T}, \mathbf{Y}_{1:T-1}|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t})] \\
= &\log Z_\theta(\mathbf{X}_{1:t}, \mathbf{Y}_{1:t-1})
\end{aligned}$$

Thus setting $C_{\mathbf{X}_{1:t},\mathbf{Y}_{1:t-1}}$ to negate this remaining term, which is also only a function of $\mathbf{X}_{1:t}$ and $\mathbf{Y}_{1:t-1}$ (and, importantly, not Y_t), completes the proof. □

Lemma A.1. $H(\boldsymbol{Y}^T||\boldsymbol{X}^T) = -\mathrm{E}_{\mathrm{P}(\boldsymbol{X},\boldsymbol{Y})}[\theta^\top \mathcal{F}(\boldsymbol{X}, \boldsymbol{Y})] + \log Z$ *where* $\log Z = \sum_{X_1} \log Z_{X_1}$

Proof.

$$
\begin{aligned}
H(\mathbf{Y}^T||\mathbf{X}^T) &= \sum_{t=1}^{T} H(Y_t|\mathbf{X}^{1:t}, \mathbf{Y}^{1:t-1}) \\
&= \sum_{t=1}^{T} \mathrm{E}_{\mathrm{P}(\mathbf{X}_{1:t},\mathbf{Y}_{1:t})} \left[-\log \mathrm{P}(Y_t|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t-1})\right] \\
&= \sum_{t=1}^{T-1} \mathrm{E}_{\mathrm{P}(\mathbf{X}_{1:t},\mathbf{Y}_{1:t})} \left[- \sum_{X_{t+1}} \mathrm{P}(X_{t+1}|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t}) \log \sum_{Y_{t+1}} Z_{Y_{t+1}|\mathbf{X}_{1:t+1},\mathbf{Y}_{\mathbf{1:t}}} + \log \sum_{Y_t} Z_{Y_t|\mathbf{X}_{1:t},\mathbf{Y}_{1:t-1}} \right] \\
&\quad + \mathrm{E}_{\mathrm{P}(\mathbf{X}_{1:T},\mathbf{Y}_{1:T})} \left[-\mathcal{F}(\mathbf{X}, \mathbf{Y}) + \log \sum_{Y_T} Z_{Y_t|\mathbf{X}_{1:t},\mathbf{Y}_{1:t-1}} \right] \\
&= \sum_{t=1}^{T-1} \mathrm{E}_{\mathrm{P}(\mathbf{X}_{1:t+1},\mathbf{Y}_{1:t})} \left[- \log \sum_{Y_{t+1}} Z_{Y_{t+1}|\mathbf{X}_{1:t+1},\mathbf{Y}_{\mathbf{1:t}}} \right] + \mathrm{E}_{\mathrm{P}(\mathbf{X}_{1:t},\mathbf{Y}_{1:t-1})} \left[\log \sum_{Y_t} Z_{Y_t|\mathbf{X}_{1:t},\mathbf{Y}_{1:t-1}} \right] \\
&\quad + \mathrm{E}_{\mathrm{P}(\mathbf{X}_{1:T},\mathbf{Y}_{1:T})} \left[-\theta^\top \mathcal{F}(\mathbf{X}, \mathbf{Y}) \right] + \mathrm{E}_{\mathrm{P}(\mathbf{X}_{1:T},\mathbf{Y}_{1:T-1})} \left[\log \sum_{Y_T} Z_{Y_T|\mathbf{X}_{1:T},\mathbf{Y}_{1:T-1}} \right] \\
&= - \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})} \left[\theta^\top \mathcal{F}(\mathbf{X}, \mathbf{Y}) \right] + \log Z
\end{aligned}
$$

□

Lemma A.2. $\nabla_\theta \log Z = \mathrm{E}_{\mathrm{P}(\boldsymbol{X},\boldsymbol{Y})}[\mathcal{F}(\boldsymbol{X}, \boldsymbol{Y})]$.

Proof. This is the special case of the more general conditional result:

$$
\begin{aligned}
\nabla_\theta \log Z(Y_t|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t-1}) &= \sum_{X_{t+1}} \mathrm{P}(X_{t+1}|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t}) \nabla_\theta \log Z(\mathbf{X}_{1:t+1}, \mathbf{Y}_{1:t}) \\
&= \sum_{X_{t+1}} \mathrm{P}(X_{t+1}|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t}) \frac{\sum_{Y_{t+1}} Z_\theta(Y_{t+1}|\mathbf{X}_{1:t+1}, \mathbf{Y}_{1:t}) \nabla_\theta \log Z_\theta(Y_t|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t-1})}{Z_\theta(\mathbf{X}_{1:t+1}, \mathbf{Y}_{1:t})} \\
&= \sum_{X_{t+1}} \left(\mathrm{P}(X_{t+1}|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t}) \sum_{Y_{t+1}} \mathrm{P}(Y_{t+1}|\mathbf{X}_{1:t+1}, \mathbf{Y}_{1:t}) \nabla_\theta \log Z_\theta(Y_t|\mathbf{X}_{1:t+1}, \mathbf{Y}_{1:t}) \right) \\
&= \mathrm{E}_{\mathrm{P}(X_{t+1},Y_{t+1})} \left[\nabla_\theta \log Z_\theta(Y_{t+1}|\mathbf{X}_{1:t+1}, \mathbf{Y}_{1:t}) | \mathbf{X}_{1:t}, \mathbf{Y}_{1:t} \right] = \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})}[\mathcal{F}(\mathbf{X}, \mathbf{Y})|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t}]
\end{aligned}
$$

The final step can be seen by "pushing" the recursion to the final time-step base case. □

Lemma A.3. *The gradient of the dual of the causally conditioned entropy optimization is* $\left(\mathrm{E}_{\mathrm{P}(\boldsymbol{X},\boldsymbol{Y})}[\mathcal{F}(\boldsymbol{X}, \boldsymbol{Y})] - \mathrm{E}_{\tilde{\mathrm{P}}(\boldsymbol{X},\boldsymbol{Y})}[\mathcal{F}(\boldsymbol{X}, \boldsymbol{Y})] \right)$*, which is the difference between the expected feature vector under the probabilistic model and the empirical feature vector given the complete policy,* $\{\mathrm{P}(Y_t|\boldsymbol{X}_{1:t}, \boldsymbol{Y}_{1:t-1})\}$.

Proof. Starting with the Lagrangian, we apply Lemma A.1 then after differentiation apply Lemma A.2.

$$\begin{aligned}
\Lambda(\theta) &= -\operatorname{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})}[\theta^\top \mathcal{F}(\mathbf{X},\mathbf{Y})] + \log Z - \theta^\top \left(\operatorname{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})}[\mathcal{F}(\mathbf{X},\mathbf{Y})] - \operatorname{E}_{\tilde{\mathrm{P}}(\mathbf{X},\mathbf{Y})}[\mathcal{F}(\mathbf{X},\mathbf{Y})]\right) \\
&= \log Z - \theta^\top \operatorname{E}_{\tilde{\mathrm{P}}(\mathbf{X},\mathbf{Y})}[\mathcal{F}(\mathbf{X},\mathbf{Y})] \\
\nabla_\theta \Lambda(\theta) &= \nabla_\theta \log Z - \operatorname{E}_{\tilde{\mathrm{P}}(\mathbf{X},\mathbf{Y})}[\mathcal{F}(\mathbf{X},\mathbf{Y})] \\
&= \operatorname{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})}[\mathcal{F}(\mathbf{X},\mathbf{Y})] - \operatorname{E}_{\tilde{\mathrm{P}}(\mathbf{X},\mathbf{Y})}[\mathcal{F}(\mathbf{X},\mathbf{Y})]
\end{aligned}$$

□

Theorem 6.4. *Maximizing the causal entropy, $H(\boldsymbol{A}^T||\boldsymbol{S}^T)$ while constrained to match (in expectation) empirical feature functions, $\operatorname{E}_{\tilde{\mathrm{P}}(\boldsymbol{S},\boldsymbol{A})}[\mathcal{F}(\boldsymbol{S},\boldsymbol{A})]$, is equivalent to maximum causal likelihood estimation of θ given data set $\{\tilde{\boldsymbol{A}},\tilde{\boldsymbol{S}}\}$ under the conditional probability distribution of Equation 6.3:*

$$\begin{aligned}
\hat{\theta} &= \operatorname*{argmax}_\theta \log \prod_i \mathrm{P}_\theta(\tilde{\boldsymbol{A}}^{(i)}||\tilde{\boldsymbol{S}}^{(i)}) \\
&= \operatorname*{argmax}_\theta \sum_{i,t} \log \mathrm{P}(\tilde{A}_t^{(i)}|\tilde{\mathbf{S}}_{1:t}^{(i)}, \tilde{\mathbf{A}}_{1:t-1}^{(i)}),
\end{aligned} \tag{A.9}$$

where (i) indexes the training examples.

Proof. We re-express Equation A.9 as follows:

$$\begin{aligned}
\sum_{i,t} \log \mathrm{P}(\tilde{A}_t^{(i)}|\tilde{\mathbf{S}}_{1:t}^{(i)}, \tilde{\mathbf{A}}_{1:t-1}^{(i)}) &= \sum_t \sum_{\mathbf{A}_{1:t},\mathbf{S}_{1:t}} \tilde{\mathrm{P}}(\mathbf{S}_{1:t},\mathbf{A}_{1:t}) \log \mathrm{P}(A_t|\mathbf{S}_{1:t},\mathbf{A}_{1:t-1}) \\
&= \sum_t \sum_{\mathbf{A}_{1:t},\mathbf{S}_{1:t}} \tilde{\mathrm{P}}(\mathbf{S}_{1:t},\mathbf{A}_{1:t}) \left(\log Z_{A_t|\mathbf{S}_{1:t},\mathbf{A}_{1:t-1}} - \log Z_{\mathbf{S}_{1:t},\mathbf{A}_{1:t-1}}\right),
\end{aligned}$$

and then find the gradient via Lemma A.2:

$$\begin{aligned}
&\nabla_\theta \left(\sum_t \sum_{\mathbf{A}_{1:t},\mathbf{S}_{1:t}} \tilde{\mathrm{P}}(\mathbf{S}_{1:t},\mathbf{A}_{1:t}) \left(\log Z_{A_t|\mathbf{S}_{1:t},\mathbf{A}_{1:t-1}} - \log Z_{\mathbf{S}_{1:t},\mathbf{A}_{1:t-1}}\right) \right) \\
&= \sum_t \sum_{\mathbf{A}_{1:t},\mathbf{S}_{1:t}} \tilde{\mathrm{P}}(\mathbf{S}_{1:t},\mathbf{A}_{1:t}) \left(\operatorname{E}_{\mathrm{P}(\mathbf{S},\mathbf{A})}\left[\mathcal{F}(\mathbf{A},\mathbf{S})|\mathbf{S}_{1:t},\mathbf{A}_{1:t}\right] - \operatorname{E}_{\mathrm{P}(\mathbf{S},\mathbf{A})}\left[\mathcal{F}(\mathbf{A},\mathbf{S})|\mathbf{S}_{1:t},\mathbf{A}_{1:t-1}\right]\right) \\
&= \operatorname{E}_{\tilde{\mathrm{P}}(\mathbf{A},\mathbf{S})}[\mathcal{F}(\mathbf{A},\mathbf{S})] - \operatorname{E}_{\mathrm{P}(\mathbf{A},\mathbf{S})}[\mathcal{F}(\mathbf{A},\mathbf{S})]
\end{aligned}$$

This gradient is equivalent (modulo the sign) to the gradient of the dual (Lemma A.3). Thus, the optimizing either function produces the same optima solution point. □

Theorem 6.6. *Any distribution that matches feature function expectations,* $\mathrm{E}_{\mathrm{P}(\boldsymbol{S},\boldsymbol{A})}[\mathcal{F}(\boldsymbol{S},\boldsymbol{A})]$ *with demonstrated expectations,* $\mathrm{E}_{\tilde{\mathrm{P}}(\boldsymbol{S},\boldsymbol{A})}[\mathcal{F}(\boldsymbol{S},\boldsymbol{A})]$*, guarantees equivalent expected utility on the unknown parameters of a reward function linear in* $\mathcal{F}(\boldsymbol{S},\boldsymbol{A})$.

Proof. Following Abbeel & Ng (2004) and simply employing the definition of the expected reward we have:

$$\begin{aligned}
\mathrm{E}_{\mathrm{P}(\mathbf{S},\mathbf{A})}[\mathcal{F}(\mathbf{S},\mathbf{A})] &= \mathrm{E}_{\tilde{\mathrm{P}}(\mathbf{S},\mathbf{A})}[\mathcal{F}(\mathbf{S},\mathbf{A})] \\
\forall_\theta \;\; \theta^\top \mathrm{E}_{\mathrm{P}(\mathbf{S},\mathbf{A})}[\mathcal{F}(\mathbf{S},\mathbf{A})] &= \theta^\top \mathrm{E}_{\tilde{\mathrm{P}}(\mathbf{S},\mathbf{A})}[\mathcal{F}(\mathbf{S},\mathbf{A})] \\
\forall_\theta \;\; \mathrm{E}_{\mathrm{P}(\mathbf{S},\mathbf{A})}[\mathrm{reward}_\theta(\mathbf{S},\mathbf{A})] &= \mathrm{E}_{\tilde{\mathrm{P}}(\mathbf{S},\mathbf{A})}[\mathrm{reward}_\theta(\mathbf{S},\mathbf{A})].
\end{aligned}$$

□

Theorem 6.8. *The maximum causal entropy distribution with statistic matching (Theorem 6.2) can be re-expressed as:*

$$\begin{aligned}
Q_\theta^{\mathit{soft}}(a_t, s_t) &\triangleq \log Z_{a_t|s_t} \\
&= \mathrm{E}_{\mathrm{P}(s_{t+1}|s_t,a_t)}[V_\theta^{\mathit{soft}}(s_{t+1})|s_t, a_t] + \theta^\top \boldsymbol{f}_{s_t,a_t} \qquad \text{(A.10)} \\
V_\theta^{\mathit{soft}}(s_t) &\triangleq \log Z_{s_t} \\
&= \operatorname*{softmax}_{a_t} Q_\theta^{\mathit{soft}}(a_t, s_t), \qquad \text{(A.11)}
\end{aligned}$$

where $\operatorname{softmax}_x f(x) \triangleq \log \sum_x e^{f(x)}$ *provides a smooth interpolation (i.e., differentiable) of the maximum of different functions.*

Proof. These relationships can be directly verified algebraically. □

Theorem 6.10. *The probability of a stochastic policy,* $\pi \triangleq \{\mathrm{P}(A_\tau|\boldsymbol{S}_{1:\tau}, \boldsymbol{A}_{1:\tau-1})\}$*, under the maximum causal entropy distribution is related to the expected feature potentials and the softmax recurrence as follows:*

$$\log \mathrm{P}_\theta^{\mathit{soft}}(\pi) = \mathrm{E}_{\pi(\boldsymbol{S}_{1:T},\boldsymbol{A}_{1:T})}\left[\sum_{t=1}^{T} \theta^\top \boldsymbol{f}_{S_t,A_t}\right] - \sum_{S_1} p(S_1) V_\theta^{\mathit{soft}}(S_1), \qquad \text{(A.12)}$$

where the latter term is independent from the policy, π.

Proof. We start from the $\log$ of the multinomial distribution and employ the soft value iteration

interpretation of the action probability distribution to prove the theorem.

$$
\begin{aligned}
\log \mathrm{P}_{\theta}^{\text{soft}}(\pi) &= \log \prod_{\mathbf{A}_{1:T},\mathbf{S}_{1:T}} \mathrm{P}(\mathbf{A}_{1:T}||\mathbf{S}_{1:T})^{\pi(\mathbf{A}_{1:T},\mathbf{S}_{1:T})} \\
&= \sum_{\mathbf{A}_{1:T},\mathbf{S}_{1:T}} \pi(\mathbf{A}_{1:T},\mathbf{S}_{1:T}) \log \mathrm{P}(\mathbf{A}^{1:T}||\mathbf{S}^{1:T}) \\
&= \sum_{t=1}^{T} \sum_{A_t,S_t} \pi(A_t,S_t) \log \mathrm{P}(A_t|S_t) \\
&= \sum_{t=1}^{T} \sum_{A_t,S_t} \pi(A_t,S_t) \left(Q_{\theta}^{\text{soft}}(A_t,S_t) - V_{\theta}^{\text{soft}}(S_t)\right) \\
&= \sum_{t=1}^{T} \sum_{A_t,S_t} \pi(A_t,S_t) \left(\sum_{S_{t+1}} \left(\mathrm{P}(S_{t+1}|S_t,A_t) V_{\theta}^{\text{soft}}(S_{t+1})\right) + \theta^{\top}\mathbf{f}_{S_t,A_t} - V_{\theta}^{\text{soft}}(S_t) \right) \\
&= \sum_{t=1}^{T} \sum_{A_t,S_t} \pi(S_t,A_t)\theta^{\top}\mathbf{f}_{S_t,A_t} + \sum_{t=2}^{T} \sum_{S_t} \pi(S_t) V_{\theta}^{\text{soft}}(S_t) - \sum_{t=1}^{T} \sum_{S_t} \pi(S_t) V_{\theta}^{\text{soft}}(S_t) \\
&= \mathrm{E}_{\pi(\mathbf{S}_{1:T},\mathbf{A}_{1:T})} \left[\sum_{t=1}^{T} \theta^{\top}\mathbf{f}_{S_t,A_t} \right] - \sum_{S_1} p(S_1) V_{\theta}^{\text{soft}}(S_1),
\end{aligned}
\tag{A.13}
$$

where $\pi(\mathbf{A}_{1:T},\mathbf{S}_{1:T})$, $\pi(A_t,S_t)$, and $\pi(S_t)$ are the probability of state-action sequences and marginal probabilities of time-indexed state-action pairs, and states. □

Theorem 6.13. *The deviation between the empirical average of feature vectors and the expectation of feature vectors is bounded by:*

$$
\mathrm{P}\left(\left|\left| \frac{1}{n}\sum_i \boldsymbol{F}_i - \mathrm{E}\left[\frac{1}{n}\sum_i \boldsymbol{F}_i\right] \right|\right|_{\infty} \geq \epsilon \right) \leq \sum_{k=1}^{K} 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(f_{i,k}^{max} - f_{i,k}^{min})^2} \right), \tag{A.14}
$$

where $\boldsymbol{F}_1, \boldsymbol{F}_2, ...$ are random variables corresponding to expected feature vectors obtained by policies (random variables), and assuming that those feature samples are bounded by $P(F_{i,k} - E[F_{i,k}] \in [f_{i,k}^{min}, f_{i,k}^{max}]) = 1$. In the special case that all elements of the difference of sampled feature vectors from their expectation are bounded by the same values, this reduces to:

$$
\mathrm{P}\left(\left|\left| \frac{1}{n}\sum_i \boldsymbol{F}_i - \mathrm{E}\left[\frac{1}{n}\sum_i \boldsymbol{F}_i\right] \right|\right|_{\infty} \geq \epsilon \right) \leq 2K\exp\left(-\frac{2n\epsilon^2}{(f^{max} - f^{min})^2} \right). \tag{A.15}
$$

Proof. By Hoeffding's inequality, we have:

$$\mathrm{P}\left(\left|\frac{1}{n}\sum_i F_{i,k} - \mathrm{E}\left[\frac{1}{n}\sum_i F_{i,k}\right]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^n (f_{i,k}^{\max} - f_{i,k}^{\min})^2}\right), \tag{A.16}$$

By the union bound:

$$\mathrm{P}\left(\bigcup_{k=1}^K \left|\frac{1}{n}\sum_i F_{i,k} - \mathrm{E}\left[\frac{1}{n}\sum_i F_{i,k}\right]\right| \geq \epsilon\right) \leq \sum_{k=1}^K \mathrm{P}\left(\left|\frac{1}{n}\sum_i F_{i,k} - \mathrm{E}\left[\frac{1}{n}\sum_i F_{i,k}\right]\right| \geq \epsilon\right).$$

Combining these, and recognizing that:

$$\mathrm{P}\left(\bigcup_{k=1}^K \left|\frac{1}{n}\sum_i F_{i,k} - \mathrm{E}\left[\frac{1}{n}\sum_i F_{i,k}\right]\right| \geq \epsilon\right) = \mathrm{P}\left(\left|\left|\frac{1}{n}\sum_i F_{i,k} - \mathrm{E}\left[\frac{1}{n}\sum_i F_{i,k}\right]\right|\right|_\infty \geq \epsilon\right)$$

proves the theorem. □

Theorem 6.14. *The deviation between the empirical average reward and the expected reward is bounded by:*

$$\mathrm{P}\left(\left|\frac{1}{n}\sum_i \theta^\top \boldsymbol{F}_i - \mathrm{E}\left[\frac{1}{n}\sum_i \theta^\top \boldsymbol{F}_i\right]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^n (\mathit{reward}_i^{\mathit{max}} - \mathit{reward}_i^{\mathit{min}})^2}\right), \tag{A.17}$$

where $\theta^\top \boldsymbol{F}_1, \theta^\top \boldsymbol{F}_2, ...,$ are the expected rewards obtained from policies with expected features $\boldsymbol{F}_1, ...$ under policies (random variables), and assuming that the rewards are bounded by $P(\theta^\top \boldsymbol{F}_i - E[\theta^\top \boldsymbol{F}_i] \in [\mathit{reward}_i^{\mathit{min}}, \mathit{reward}_i^{\mathit{max}}]) = 1$. In the special case that the bounds on the rewards are the same, this reduces to:

$$\mathrm{P}\left(\left|\frac{1}{n}\sum_i \theta^\top \boldsymbol{F}_i - \mathrm{E}\left[\frac{1}{n}\sum_i \theta^\top \boldsymbol{F}_i\right]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2n\epsilon^2}{(\mathit{reward}^{\mathit{max}} - \mathit{reward}^{\mathit{min}})^2}\right). \tag{A.18}$$

Proof. A direct application of Hoeffding's inequality proves the theorem. □

Theorem 6.15. *For the special case where dynamics are linear functions with Gaussian noise, the quadratic MaxCausalEnt model permits a closed-form solution and, given dynamics $\boldsymbol{s}_{t+1} \sim N(\boldsymbol{A}\boldsymbol{s}_t + \boldsymbol{B}\boldsymbol{a}_t, \Sigma)$, Equation 2.10 reduces to:*

$$Q_\theta^{\mathit{soft}}(\boldsymbol{a}_t, \boldsymbol{s}_t) = \begin{bmatrix} \boldsymbol{a}_t \\ \boldsymbol{s}_t \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{B}^\top \boldsymbol{D}\boldsymbol{B} & \boldsymbol{A}^\top \boldsymbol{D}\boldsymbol{B} \\ \boldsymbol{B}^\top \boldsymbol{D}\boldsymbol{A} & \boldsymbol{A}^\top \boldsymbol{D}\boldsymbol{A} \end{bmatrix} \begin{bmatrix} \boldsymbol{a}_t \\ \boldsymbol{s}_t \end{bmatrix} + \begin{bmatrix} \boldsymbol{a}_t \\ \boldsymbol{s}_t \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{B}^\top \boldsymbol{G} \\ \boldsymbol{A}^\top \boldsymbol{G} \end{bmatrix}$$

$$V_\theta^{\mathit{soft}}(\boldsymbol{s}_t) = \boldsymbol{s}_t^\top (\boldsymbol{C}_{s,s} + Q - \boldsymbol{C}_{a,s}^\top \boldsymbol{C}_{a,a}^{-1} \boldsymbol{C}_{a,s}) \boldsymbol{s}_t + \boldsymbol{s}_t^\top (\boldsymbol{F}_s + \boldsymbol{R}) + \mathit{const},$$

where $\boldsymbol{C}$ *and* $\boldsymbol{D}$ *are recursively computed as:* $\boldsymbol{C}_{a,a} = \boldsymbol{B}^\top \boldsymbol{D}\boldsymbol{B}; \boldsymbol{C}_{s,a} = \boldsymbol{C}_{a,s}^\top = \boldsymbol{B}^\top \boldsymbol{D}\boldsymbol{A}; \boldsymbol{C}_{s,s} = \boldsymbol{A}^\top \boldsymbol{D}\boldsymbol{A}; \boldsymbol{D} = \boldsymbol{C}_{s,s} + \boldsymbol{Q} - \boldsymbol{C}^\top \boldsymbol{C}_{a,a}^{-1} \boldsymbol{C}_{a,s}$*; and* $\boldsymbol{G} = \boldsymbol{F}_s + \boldsymbol{R}$*.*

Proof. Maximizing the causally conditioned entropy, $H(\mathbf{a}||\mathbf{s})$, subject to quadratic constraints yields a temporal recurrence over the continuous state and action spaces (Theorem 6.2). Using the $\mathbf{Q}$ and $\mathbf{R}$ matrices as Lagrange multipliers, we arrive at the quadratic LQR evaluation metric, $\sum_t tr[\mathbf{a}_t \mathbf{a}_t^\top \mathbf{Q}] = \sum_t \mathbf{a}_t^\top \mathbf{Q} \mathbf{a}_t$ and $\sum_t tr[\mathbf{s}_t \mathbf{s}_t^\top \mathbf{R}] = \sum_t \mathbf{s}_t^\top \mathbf{R} \mathbf{s}_t$, and the dynamic programming algorithm for recursive computation is:

$$Q_\theta^{\text{soft}}(\mathbf{a}_t, \mathbf{s}_t) = \int_{\mathbf{s}_{t+1}} \mathrm{P}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) V_\theta^{\text{soft}}(\mathbf{s}_{t+1}) d\mathbf{s}_{t+1} \tag{A.19}$$

$$V_\theta^{\text{soft}}(\mathbf{s}_t) = \operatorname*{softmax}_{\mathbf{a}_t} Q_\theta^{\text{soft}}(\mathbf{a}_t, \mathbf{s}_t) + \mathbf{s}_t^\top \mathbf{Q} \mathbf{s}_t + \mathbf{s}_t^\top \mathbf{R}. \tag{A.20}$$

We will assume and then verify that both $Q_\theta^{\text{soft}}(\mathbf{a}_t, \mathbf{s}_t)$ and $V_\theta^{\text{soft}}(\mathbf{s}_t)$ have quadratic forms; specifically,

$$Q_\theta^{\text{soft}}(\mathbf{a}_t, \mathbf{s}_t) = \begin{bmatrix} \mathbf{a}_t \\ \mathbf{s}_t \end{bmatrix}^\top \begin{bmatrix} \mathbf{C}_{a,a} & \mathbf{C}_{a,s} \\ \mathbf{C}_{s,a} & \mathbf{C}_{s,s} \end{bmatrix} \begin{bmatrix} \mathbf{a}_t \\ \mathbf{s}_t \end{bmatrix} + \begin{bmatrix} \mathbf{a}_t \\ \mathbf{s}_t \end{bmatrix}^\top \begin{bmatrix} \mathbf{F}_a \\ \mathbf{F}_s \end{bmatrix}, \text{ and}$$
$$V_\theta^{\text{soft}}(\mathbf{s}_t) = \mathbf{s}_t^\top \mathbf{D} \mathbf{s}_t + \mathbf{s}_t^\top \mathbf{G}.$$

Based on this assumption, Equation A.19 can be equivalently expressed as:

$$\begin{aligned} Q_\theta^{\text{soft}}(\mathbf{a}_t, \mathbf{s}_t) &= \mathrm{E}_{\mathbf{s}_{t+1}}[\mathbf{s}_{t+1}^\top \mathbf{D} \mathbf{s}_{t+1} + \mathbf{s}_{t+1}^\top \mathbf{G} | \mathbf{s}_t, \mathbf{a}_t] \\ &= (\mathbf{A}\mathbf{s}_t + \mathbf{B}\mathbf{a}_t)^\top \mathbf{D} (\mathbf{A}\mathbf{s}_t + \mathbf{B}\mathbf{a}_t) + \mathrm{tr}(\mathbf{D}\Sigma) + \mathbf{a}_t^\top \mathbf{B}^\top \mathbf{G} + \mathbf{s}_t^\top \mathbf{A}^\top \mathbf{G} \\ &= \begin{bmatrix} \mathbf{a}_t \\ \mathbf{s}_t \end{bmatrix}^\top \begin{bmatrix} \mathbf{B}^\top \mathbf{D} \mathbf{B} & \mathbf{A}^\top \mathbf{D} \mathbf{B} \\ \mathbf{B}^\top \mathbf{D} \mathbf{A} & \mathbf{A}^\top \mathbf{D} \mathbf{A} \end{bmatrix} \begin{bmatrix} \mathbf{a}_t \\ \mathbf{s}_t \end{bmatrix} + \begin{bmatrix} \mathbf{a}_t \\ \mathbf{s}_t \end{bmatrix}^\top \begin{bmatrix} \mathbf{B}^\top \mathbf{G} \\ \mathbf{A}^\top \mathbf{G} \end{bmatrix} + \text{const.} \end{aligned} \tag{A.21}$$

Thus the set of update rules are: $\mathbf{C}_{a,a} = \mathbf{B}^\top \mathbf{D}\mathbf{B}$; $\mathbf{C}_{s,a} = \mathbf{C}_{a,s}^\top = \mathbf{B}^\top \mathbf{D}\mathbf{A}$; $\mathbf{C}_{s,s} = \mathbf{A}^\top \mathbf{D}\mathbf{A}$; $\mathbf{F}_a = \mathbf{B}^\top \mathbf{G}$; and $\mathbf{F}_s = \mathbf{A}^\top \mathbf{G}$.

Equation A.20 can be expressed as:

$$\begin{aligned} V_\theta^{\text{soft}}(\mathbf{s}_t) &= \log \int_{\mathbf{a}_t} e^{\mathbf{a}_t^\top \mathbf{C}_{a,a} \mathbf{a}_t + 2\mathbf{a}_t^\top \mathbf{C}_{a,s} \mathbf{s}_t + \mathbf{s}_t \mathbf{C}_{s,s} \mathbf{s}_t + \mathbf{s}_t^\top \mathbf{F}_s + \mathbf{a}_t^\top \mathbf{F}_a} d\mathbf{a}_t + \mathbf{s}_t^\top \mathbf{Q} \mathbf{s}_t + \mathbf{s}_t^\top \mathbf{R} \\ &= \log \int_{\mathbf{a}_t} e^{(\mathbf{a}_t + \mathbf{C}_{a,a}^{-1} \mathbf{C}_{a,s} \mathbf{s}_t)^\top \mathbf{C}_{a,a} (\mathbf{a}_t + \mathbf{C}_{a,a}^{-1} \mathbf{C}_{a,s} \mathbf{s}_t) - \mathbf{s}_t^\top \mathbf{C}_{a,s}^\top \mathbf{C}_{a,a}^{-1\top} \mathbf{C}_{a,a} \mathbf{C}_{a,a}^{-1} \mathbf{C}_{a,s} \mathbf{s}_t + \mathbf{s}_t^\top \mathbf{C}_{s,s} \mathbf{s}_t + \mathbf{s}_t^\top \mathbf{F}_s + \mathbf{a}_t^\top \mathbf{F}_a} d\mathbf{a}_t \\ &\quad + \mathbf{s}_t^\top \mathbf{Q} \mathbf{s}_t + \mathbf{s}_t^\top \mathbf{R} \\ &= \log \sqrt{\pi^d |\mathbf{C}_{a,a}|} - \mathbf{s}_t^\top \mathbf{C}_{a,s}^\top \mathbf{C}_{a,a}^{-1} \mathbf{C}_{a,s} \mathbf{s}_t + \mathbf{s}_t^\top \mathbf{C}_{s,s} \mathbf{s}_t + \mathbf{s}_t^\top \mathbf{F}_s + \mathbf{s}_t^\top \mathbf{Q} \mathbf{s}_t + \mathbf{s}_t^\top \mathbf{R} \\ &= \mathbf{s}_t^\top (\mathbf{C}_{s,s} + \mathbf{Q} - \mathbf{C}_{a,s}^\top \mathbf{C}_{a,a}^{-1} \mathbf{C}_{a,s}) \mathbf{s}_t + \mathbf{s}_t^\top (\mathbf{F}_s + \mathbf{R}) + \text{const}, \end{aligned}$$

yielding update rule: $\mathbf{D} = \mathbf{C}_{s,s} + \mathbf{Q} - \mathbf{C}_{a,s}^\top \mathbf{C}_{a,a}^{-1} \mathbf{C}_{a,s}$ and $\mathbf{G} = \mathbf{F}_s + \mathbf{R}$. The probabilistic policy under this model is: $\pi(\mathbf{a}_t|\mathbf{s}_t) \propto e^{Q_\theta^{\text{soft}}(\mathbf{a}_t, \mathbf{s}_t)}$. □

Theorem 6.16. *The marginalized maximum conditional entropy distribution (Equation 6.23) can be interpreted under the dynamic programming perspective using the re-expression:*

$$Q_\theta^{CE}(a_t, s_t) = \operatorname*{softmax}_{s_{t+1}} \left(V_\theta^{CE}(s_{t+1}) + \log \mathrm{P}(s_{t+1}|s_t, a_t) + \theta^\top \boldsymbol{f}_{a_t, s_t} \right) \tag{A.22}$$
$$V_\theta^{CE}(s_t) = \operatorname*{softmax}_{a_t} Q_\theta^{CE}(a_t, s_t)$$

Proof. The conditional probability of Y_t is distributed as:

$$\begin{aligned} \mathrm{P}(Y_t|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t-1}) &\propto \sum_{\mathbf{X}_{t+1:T}, \mathbf{Y}_{t+1:T}} \left(\prod_{\tau=t}^{T} \mathrm{P}(X_{\tau+1}|X_\tau, Y_\tau) \right) e^{\theta^\top \mathcal{F}(\mathbf{X}, \mathbf{Y})} \\ &= \sum_{\mathbf{X}_{t+1:T}, \mathbf{Y}_{t+1:T}} e^{\sum_{\tau=t}^{T} \theta^\top \mathcal{F}(\mathbf{X}, \mathbf{Y}) + \log \mathrm{P}(X_{\tau+1}|X_\tau, Y_\tau)} \end{aligned} \tag{A.23}$$

Note that $\mathrm{P}(Y_t|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t-1}) = \mathrm{P}(Y_t|X_t)$ given the factorization of $\mathcal{F}$ and the side information distribution. We define $Q_\theta^{\text{lcrf}}(Y_t, X_t)$ as the log of this unnormalized probability (Equation A.23) and will now recursively define it, eventually in terms of a $V_\theta^{\text{lcrf}}(X_t)$ term, which we will also define.

$$\begin{aligned} Q_\theta^{\text{lcrf}}(Y_t, X_t) &= \log \sum_{\mathbf{X}_{t+1:T}, \mathbf{Y}_{t+1:T}} e^{\sum_{\tau=t+1}^{T} \theta^\top \mathcal{F}(\mathbf{X}, \mathbf{Y}) + \log \mathrm{P}(X_{\tau+1}|X_\tau, Y_\tau)} \\ &= \log \sum_{X_{t+1}, Y_{t+1}} \left(e^{\theta^\top \mathcal{F}_{X_{t+1}} + \log \mathrm{P}(X_{t+1}|X_t, Y_t)} \sum_{\mathbf{X}_{t+2:T}, \mathbf{Y}_{t+2:T}} e^{\sum_{\tau=t+1}^{T} \theta^\top \mathcal{F}_{X_{\tau+1}} + \log \mathrm{P}(X_{\tau+1}|X_\tau, Y_\tau)} \right) \\ &= \log \sum_{X_{t+1}, Y_{t+1}} \left(e^{\theta^\top \mathcal{F}_{X_{t+1}} + \log \mathrm{P}(X_{t+1}|X_t, Y_t) + Q_\theta^{\text{lcrf}}(Y_{t+1}, X_{t+1})} \right) \\ &= \log \sum_{X_{t+1}} \left(e^{\log \mathrm{P}(X_{t+1}|X_t, Y_t) + V_\theta^{\text{lcrf}}(X_{t+1})} \right) \\ &= \operatorname*{softmax}_{X_{t+1}} V_\theta^{\text{lcrf}}(X_{t+1}) + \log \mathrm{P}(X_{t+1}|X_t, Y_t) \\ V_\theta^{\text{lcrf}}(X_t) &= \log \sum_{Y_t} e^{Q_\theta^{\text{lcrf}}(Y_t, X_t) + \theta^\top \mathcal{F}_{X_t}} \\ &= \operatorname*{softmax}_{Y_t} Q_\theta^{\text{lcrf}}(Y_t, X_t) + \theta^\top \mathcal{F}_{X_t} \end{aligned}$$

□

A.4 Chapter 7 Proofs

Theorem 7.6. *The maximum causal entropy probability distribution for the imperfect information setting with perfect past decision recall (Theorem 7.6) is distributed according to the following*

recurrence relationship:

$$
\begin{aligned}
P_\theta(Y_t|\boldsymbol{par}(Y_t)) &= \frac{Z_{Y_t|\boldsymbol{par}(Y_t),\theta}}{Z_{\boldsymbol{par}(Y_t),\theta}} \\
\log Z_{\boldsymbol{par}(Y_t),\theta} &= \log \sum_{Y_t} Z_{Y_t|\boldsymbol{par}(Y_t),\theta} \\
&= \operatorname*{softmax}_{Y_t} \Big(\mathrm{E}_{\mathrm{P}(\boldsymbol{par}(Y_{t+1})|\boldsymbol{par}(Y_t),Y_t)} \left[\log Z_{\boldsymbol{par}(Y_t),\theta} | \boldsymbol{par}(Y_t), Y_t) \right] \Big) \\
Z_{Y_t|\boldsymbol{par}(Y_t),\theta} &= e^{\left(\mathrm{E}_{\mathrm{P}(\boldsymbol{par}(Y_{t+1})|\boldsymbol{par}(Y_t),Y_t)} \left[\log Z_{\boldsymbol{par}(Y_t),\theta} | \boldsymbol{par}(Y_t),Y_t) \right] \right)} \\
Z_{\boldsymbol{par}(Y_{T+1}),\theta} &= e^{\theta^\top \mathcal{F}(\boldsymbol{X},\boldsymbol{Y})},
\end{aligned}
\tag{A.24}
$$

where the final set of parents for the "after last" Y variable is the complete set of variables: $\boldsymbol{par}(Y_{T+1}) \triangleq \boldsymbol{X} \cup \boldsymbol{Y}$.

Proof. As in the proof of Theorem 5.8, we consider the complete causally conditioned sequence of variables $\mathrm{P}(\mathbf{Y}||\mathbf{X})$. The convex optimization is then:

$$
\begin{aligned}
&\operatorname*{argmin}_{\{\mathrm{P}(\mathbf{Y}||\mathbf{X})\}} -H(\mathbf{Y}||\mathrm{par}(\mathbf{Y})) \\
\text{such that: } &\mathrm{E}_{\mathbf{X},\mathbf{Y}}[f_k(\mathbf{X},\mathbf{Y})] = \tilde{\mathrm{E}}_{\mathbf{X},\mathbf{Y}}[f_k(\mathbf{X},\mathbf{Y})] \text{ for } k = 1, ..., K; \\
&\forall_{\mathrm{par}(\mathbf{Y})} \sum_{\mathbf{Y}} \mathrm{P}(\mathbf{Y}||\mathrm{par}(\mathbf{Y})) = 1; \\
\text{and: } &\forall_{t,\mathbf{X},\hat{\mathbf{X}},\mathbf{Y},\hat{\mathbf{Y}}:\{\mathbf{X},\mathbf{Y}\}_{\mathrm{par}(Y_t)}=\{\hat{\mathbf{X}},\hat{\mathbf{Y}}\}_{\mathrm{par}(Y_t)}}, \sum_{\mathbf{Y}_{t+1:T}} \mathrm{P}(\mathbf{Y}||\mathbf{X}) = \sum_{\mathbf{Y}_{t+1:T}} \mathrm{P}(\hat{\mathbf{Y}}||\hat{\mathbf{X}}),
\end{aligned}
\tag{A.25}
$$

where $\{\mathbf{X},\mathbf{Y}\}_{\mathrm{par}(Y_t)}$ indicates the $\mathbf{X}$ and $\mathbf{Y}$ variables that are parents of Y_t. The final constraint of A.25 enforces the causal constraint and constrains the conditional probabilities of Y_t with different non-parent variables to be the same. With the parent sets constrained by perfect past decision recall, these constraints are linear functions of causally conditioned probabilities, so convexity is maintained. □

Theorem 7.8. *The maximum causal entropy probability distribution for the imperfect information*

setting is distributed according to the following recurrence relationship:

$$
\begin{aligned}
\mathrm{P}_\theta(Y_t|\mathit{par}(Y_t)) &= \frac{Z_{Y_t|\mathit{par}(Y_t),\theta}}{Z_{\mathit{par}(Y_t),\theta}} \\
\log Z_{\mathit{par}(Y_t),\theta} &= \log \sum_{Y_t} Z_{Y_t|\mathit{par}(Y_t),\theta} \\
&= \operatorname*{softmax}_{Y_t} \Big(\mathrm{E}_{\mathrm{P}(\mathit{par}(Y_{t+1})|\mathit{par}(Y_t),Y_t)} \left[\log Z_{\mathit{par}(Y_t),\theta} | \mathit{par}(Y_t), Y_t \right] \Big) \\
Z_{Y_t|\mathit{par}(Y_t),\theta} &= e^{\left(\mathrm{E}_{\mathrm{P}(\mathit{par}(Y_{t+1})|\mathit{par}(Y_t),Y_t)}\left[\log Z_{\mathit{par}(Y_t),\theta}|\mathit{par}(Y_t),Y_t\right]\right)} \\
Z_{\mathit{par}(Y_{T+1}),\theta} &= e^{\theta^\top \mathcal{F}(\boldsymbol{X},\boldsymbol{Y})},
\end{aligned}
\tag{A.26}
$$

where the final set of parents for the "after last" conditioned variable, Y_{T+1}, is the complete set of variables: $\mathit{par}(Y_{T+1}) \triangleq \boldsymbol{X} \cup \boldsymbol{Y}$.

Proof. Differentiating the Lagrangian of the maximum causal entropy optimization (Equation A.25),

$$
\begin{aligned}
\Lambda(\mathrm{P}, \theta) = H(\mathbf{Y}^T || \mathit{par}(\mathbf{Y})^T) &+ \sum_k \theta_k \left(\mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})}[\mathcal{F}_k(\mathbf{X},\mathbf{Y})] - \mathrm{E}_{\tilde{\mathrm{P}}(\mathbf{X},\mathbf{Y})}[\mathcal{F}_k(\mathbf{X},\mathbf{Y})] \right) \\
&+ \sum_{t, \mathbf{X}^{1:t}, \mathbf{Y}^{1:t-1}} C_{\mathrm{par}(Y_t)} \left(\sum_{Y_t} \mathrm{P}(Y_t|\mathrm{par}(Y_t)) - 1 \right)
\end{aligned}
\tag{A.27}
$$

we have:

$$
\begin{aligned}
\nabla_{\{\mathrm{P}(Y_t|\mathit{par}(Y_t))\}} \Lambda(\mathrm{P}, \theta) = \Bigg\{ & C_{\mathrm{par}(Y_t)} - \mathrm{P}_\theta(\mathrm{par}(Y_t)) \\
& \left(\sum_{\tau=t}^{T} \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})} \left[\log \mathrm{P}_\theta(Y_\tau|\mathrm{par}(Y_\tau)) | \mathrm{par}(Y_t), Y_t \right] + \sum_k \theta_k \, \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})}[F_k(\mathbf{X},\mathbf{Y})|\mathrm{par}(Y_t), Y_t] \right) \Bigg\}
\end{aligned}
\tag{A.28}
$$

We ignore the $C_{\mathbf{X}_{1:t},\mathbf{Y}_{1:t-1}}$ term for now and ultimately set it to the remaining normalization term

after factoring out the $\mathrm{P}(\mathrm{par}(Y_t))$ multiplier and substituting our recursive definitions.

$$
\begin{aligned}
&- \sum_{\tau=t}^{T} \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})}\left[\log \mathrm{P}_\theta(Y_\tau|\mathrm{par}(Y_\tau))\middle|\mathrm{par}(Y_t), Y_t\right] + \theta^\top \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})}\left[\mathcal{F}(\mathbf{X},\mathbf{Y})\middle|\mathrm{par}(Y_t), Y_t\right] \\
=& - \sum_{\tau=t}^{T-1} \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})}\Big[\sum_{X_{\tau+1}} \mathrm{P}_\theta(X_{\tau+1}|\mathrm{par}(X_{\tau+1}))\log Z_\theta(\mathrm{par}(Y_{\tau+1})) - \log Z_\theta(\mathrm{par}(Y_\tau))\Big|\mathrm{par}(Y_t), Y_t\Big] \\
&- \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})}[\theta^\top \mathcal{F}(\mathbf{X},\mathbf{Y}) - \log Z_\theta(\mathrm{par}(Y_T), Y_T)|\mathrm{par}(Y_t), Y_t] + \theta^\top \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})}\left[\mathcal{F}(\mathbf{X},\mathbf{Y})\middle|\mathrm{par}(Y_t), Y_t\right] \\
=& - \sum_{\tau=t}^{T} \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})}\left[\log Z_\theta(\mathbf{X}_{1:\tau+1}, \mathbf{Y}_{1:\tau}) - \log Z_\theta(\mathbf{X}_{1:\tau}, \mathbf{Y}_{1:\tau-1})\middle|\mathrm{par}(Y_t), Y_t\right] \\
&- \mathrm{E}_{\mathrm{P}(\mathbf{X},\mathbf{Y})}[\log Z_\theta(\mathrm{par}(Y_T))|\mathrm{par}(Y_t), Y_t] \\
=& \log Z_\theta(\mathrm{par}(Y_t))
\end{aligned}
$$

Thus setting $C_{\mathrm{par}(Y_t)}$ to negate this remaining term, which is also only a function of $\mathrm{par}(Y_t)$ (and, importantly, not Y_t), completes the proof. □

Theorem 7.10. *The perfect recall maximum causal entropy distribution of Equation 7.4 can be reduced to a non-latent maximum causal entropy model by employing expectation to obtain side information dynamics,*

$$
\begin{aligned}
&\mathrm{P}\left(\boldsymbol{par}(Y_t)|Y_{t-1}, \boldsymbol{par}(Y_{t-1})\right) = \\
&\quad \mathrm{E}_{\mathrm{P}(\boldsymbol{X}|Y_{t-1},\boldsymbol{par}(Y_{t-1}))}\left[\mathrm{P}(\boldsymbol{par}(Y_t)|\boldsymbol{X}, Y_{t-1}, \boldsymbol{par}(Y_{t-1}))\ |Y_{t-1}, \boldsymbol{par}(Y_{t-1})\right],
\end{aligned}
$$

and expected statistic-based features,

$$
\mathcal{F}_t\left(Y_t, \boldsymbol{par}(Y_t)\right) = \mathrm{E}_{\mathrm{P}(\boldsymbol{X}|\boldsymbol{par}(Y_t),Y_t)}\left[\mathcal{F}_t(\boldsymbol{X}, \boldsymbol{Y}_{1:t}))\middle|\boldsymbol{par}(Y_t), Y_t\right].
$$

Proof (sketch). Distributing the feature potential function into the recursive portion of Equation A.25, we have:

$$
Z_{Y_t|\mathrm{par}(Y_t),\theta} = e^{\left(\mathrm{E}_{\mathrm{P}(\mathrm{par}(Y_{t+1})|\mathrm{par}(Y_t),Y_t)}\left[\log Z_{\mathrm{par}(Y_t),\theta}|\mathrm{par}(Y_t),Y_t\right]\right) + \mathrm{E}[\theta^\top \mathcal{F}_t(\mathbf{X},\mathbf{Y}_{1:t})|\mathrm{par}(Y_t),Y_t]}
$$

Taking logarithms and employing the transition dynamics of the theorem, the distribution can then be converted to the softmax recursive form of statistic-matching without latent side information. □

A.5 Chapter 8 Proofs

Theorem 8.8. *A LP/CP formulation of CE for Markov games is possible by considering as variables the entire sequence of joint player actions for the sequence of revealed states, $\eta(A^{1:T}_{1:N}|S^{1:T})$, and employing appropriate inequality constraints (deviation regret guarantees) and equality constraints (forcing the strategy over sequences to factor into products of Markovian strategies) on marginal distributions using linear function of $\eta(A^{1:T}_{1:N}|S^{1:T})$ variables.*

Proof. We consider optimizing over variables that represent the probability of an entire sequence of actions given the entire sequence of S states, denoted as conditional strategy sequence variables: $\eta(a^{1:T}_{1:N}|s^{1:T})$. Given the state dynamics $\mathrm{P}(s^{t+1}|s^t, a^t)$, action-state strategy probabilities, $\pi(a^t|s^{t:T})$, can be obtained by marginalizing over a linear function of conditional sequence variables:

$$\begin{aligned}\pi(a^t_{1:N}|s^{t:T}) &= \sum_{s^{1:t-1}} \sum_{a^{1:t-1}_{1:N}} \sum_{a^{t+1:T}_{1:N}} \eta(a^{1:T}_{1:N}|s^{1:T}) \prod_{\tau=1}^{t-1} \mathrm{P}(s^\tau|a^{\tau-1}, s^{\tau-1}) \\ &= \sum_{s^{1:t-1}} \sum_{a^{1:t-1}_{1:N}} \sum_{a^{t+1:T}_{1:N}} \mathrm{P}(a^{1:T}_{1:N}, s^{1:t-1}|s^{t:T}). \end{aligned} \tag{A.29}$$

Crucially, to match the Markov game setting, the conditional distribution of actions at time step t should be equivalent regardless of future state variables, $s^{t+1:T}$, since those variables are not yet known in the Markov game:

$$\forall_{t, s^{t+1:T}, \tilde{s}^{t+1:T}} \pi(a^t_{1:N}|s^t, s^{t+1:T}) = \pi(a^t_{1:N}|s^t, \tilde{s}^{t+1:T}). \tag{A.30}$$

We note that the constraints of Equation A.30 are linear functions of conditional strategy sequence variables via the steps of Equation A.29.

$$\underset{\{\eta(a_{1:N}^{1:T}|s^{1:T})\}}{\operatorname{argmax}} \quad f_0(\{\eta(a_{1:N}^{1:T}|s^{1:T})\}) \tag{A.31}$$

$$\begin{aligned}
\text{such that: } \forall_{t,i,s^t,a_i^t,a_i^{t\prime}} & \sum_{a_{-i}^t,a_{1:N}^{t+1:T},s^{1:t-1},s^{t+1:T}} \eta(a_{1:N}^{1:T}|s^{1:T}) \; \frac{\prod_{\tau=1}^{T} \mathrm{P}(s^{\tau+1}|s^\tau,a^\tau)}{\mathrm{P}(s^{t+1}|s^t,a^t)} \times \\
& \left(\mathrm{P}(s^{t+1}|s^t,a_{1:N}^{t}{}') \left(\sum_{\tau>t} \text{Utility}(s^\tau,a_{1:N}^\tau) + \text{Utility}(s^t,a_{1:N}^{t}{}') \right) \right. \\
& \left. - \mathrm{P}(s^{t+1}|s^t,a_{1:N}^t) \left(\sum_{\tau>t} \text{Utility}(s^\tau,a_{1:N}^\tau) + \text{Utility}(s^t,a_{1:N}^t) \right) \right) \leq 0 \\
\forall_{t,s^t,a_{1:N}^t} & \sum_{s^{1:t-1}} \sum_{a_{1:N}^{1:t-1}} \sum_{a_{1:N}^{t+1:T}} \eta(a_{1:N}^{1:T}|s^{1:T}) \prod_{\tau=1}^{t-1} \mathrm{P}(s^\tau|a^{\tau-1},s^{\tau-1}) \geq 0 \\
\forall_{t,s^t} & \sum_{s^{1:t-1}} \sum_{a_{1:N}^{1:t}} \sum_{a_{1:N}^{t+1:T}} \eta(a_{1:N}^{1:T}|s^{1:T}) \prod_{\tau=1}^{t-1} \mathrm{P}(s^\tau|a^{\tau-1},s^{\tau-1}) = 1 \\
\forall_{t,s^{t+1:T},\tilde{s}^{t+1:T}} & \sum_{s^{1:t-1}} \sum_{a_{1:N}^{1:t-1}} \sum_{a_{1:N}^{t+1:T}} \prod_{\tau=1}^{t-1} \mathrm{P}(s^\tau|a^{\tau-1},s^{\tau-1}) \Big(\eta\left(a_{1:N}^{1:T}|s^{1:T}\right) \\
& - \eta\left(a_{1:N}^{1:T}|s^{1:t},\tilde{s}^{t+1:T}\right) \Big) = 0
\end{aligned} \tag{A.32}$$

All constraints are linear in conditional variables, so when $-f_0$ is a linear or convex function, the optimization (Equation A.31) is a linear program or convex program. We note that the number of constraints in the last set of constraints (Equation A.32) can be reduced from $O(|S|^{2T})$ to $O(|S|^T)$ by chains of equality constraints (rather than all pair-wise constraints). However, the total number of constraints is still exponential in T and there are a total of $O(|S|^T|A|^{NT})$ variables in this formulation. □

Theorem 8.10. *Given an MCECE strategy profile, no player may decrease the predictability of her action sequence without creating deviation regret for herself.*

Proof. Ignoring all the deviation regret constraints in our notation, consider the decomposition of the causally conditioned entropy using the chain rule:

$$\begin{aligned}
\underset{\{\pi(a_{1:N}^t|s^t)\}}{\operatorname{argmax}} H(A_{1:N}^T||S^T) &= \underset{\{\pi(a_{1:N}^t|s^t)\}}{\operatorname{argmax}} \left(H(A_i^T||S^T) + H(A_{-i}^T||S^T) \right) \\
&= \left\{ \pi^{\text{MCECE}}(a_{-i}^t|s_t) \right\} \cup \underset{\{\pi(a_i^t|s^t)\}}{\operatorname{argmax}} H(A_i^T||S^T, A_{-i}^T).
\end{aligned}$$

As shown, this can equivalently be viewed as a causally conditioned entropy maximization of player i's strategy profile (with the suppressed deviation regret constraints) given the combined MCECE strategy profile of the other players. By definition this is already the least predictable strategy profile that player i can employ (subject to any deviation regret constraints). □

Theorem 8.11. *The MCECE solution strategy profile, π^{MCECE} minimizes the worst-case log prediction loss for the sequences of joint actions, i.e.,*

$$\inf_{\mathrm{P}(\boldsymbol{A}^T||\boldsymbol{S}^T)} \sup_{\tilde{\mathrm{P}}(\boldsymbol{A}^T||\boldsymbol{S}^T)} \sum_{\boldsymbol{A},\boldsymbol{S}} \tilde{\mathrm{P}}(\boldsymbol{A},\boldsymbol{S}) \log \mathrm{P}(\boldsymbol{A}^T||\boldsymbol{S}^T), \tag{A.33}$$

of all the CE satisfying deviation regret constraints, where $\tilde{\mathrm{P}}(\boldsymbol{A}^T||\boldsymbol{S}^T)$ is the (worst possible for prediction) empirical CE strategy and the joint, $\tilde{\mathrm{P}}(\boldsymbol{A},\boldsymbol{S})$, is the distribution of states and actions under that strategy profile and the known state transition dynamics.

Proof. As a special case of Grünwald & Dawid (2003), the causal entropy can be expressed as: $H(\tilde{\mathrm{P}}(\mathbf{A}^T||\mathbf{S}^T)) = \inf_{\mathrm{P}(\mathbf{A}^T||\mathbf{S}^T)} \mathrm{E}_{\tilde{\mathrm{P}}(\mathbf{A},\mathbf{S})}[-\log \mathrm{P}(\mathbf{A}^T||\mathbf{S}^T)]$. Choosing $\tilde{\mathrm{P}}(\mathbf{Y}^T||\mathbf{X}^T)$ that maximizes this then: $\sup_{\tilde{\mathrm{P}}(\mathbf{A}^T||\mathbf{S}^T)} \inf_{\mathrm{P}(\mathbf{A}^T||\mathbf{S}^T)} \mathrm{E}_{\tilde{\mathrm{P}}(\mathbf{A},\mathbf{S})}[-\log \mathrm{P}(\mathbf{A}^T||\mathbf{S}^T)]$, which is invariant to swapping the order of the $\sup$ and $\inf$ operations. □

Lemma 8.14. *The MCECE strategy profile for a Markov game is also Markovian.*

Proof sketch. The Markovian state transition dynamics of the Markov game imply that the entropy of future actions due to those state dynamics is conditionally independent of past states and actions given the current state of the game. As a consequence, any strategy profile based on certain previous states and actions can be employed from the current state by *pretending* those previous states and actions occurred, and the result is the same amount of entropy (and satisfaction of all deviation regret constraints) as if those previous states and actions had actually occurred. Thus by maximization and uniqueness of MCECE strategy profile, all MCECE strategies from the current state forward must employ the same (history-dependent) strategy profile, effectively making that strategy profile Markovian. □

Theorem 8.15. *The MCECE strategy profile, $\pi_\lambda^{MCECE}(a_{1:N}^t|s^t)$, has the following recursive form (with $\lambda \geq 0$):*

$$\pi_\lambda^{MCECE}(a_{1:N}^t|s^t) \propto e^{-\left(\sum_{i,a_i^{t\prime}} \lambda_{i,s^t,a_i^t,a_i^{t\prime}} \mathit{ExpectDevGain}_i^{\pi^{MCECE}}\left(a_{1:N}^t,s^t,a_i^{t\prime}\right)\right)+\mathit{ExpectEnt}\left(a_{1:N}^t,s^t\right)}, \tag{A.34}$$

where $\mathit{ExpectEnt}(a_{1:N}^t, s^t) \triangleq \mathrm{E}_{a_{1:N}^{t+1},s^{t+1}}\left[\mathit{ExpectEnt}(a_{1:N}^{t+1}, s^{t+1}) + H(a_{1:N}^{t+1}|s^{t+1})\middle|a_{1:N}^t, s^t\right]$.

Proof sketch. We find the form of the probability distribution by finding the optimal point of the Lagrangian. We suppress the probabilistic positivity constraints and normalization constraints with the understanding that the resulting probability distribution must normalize to 1.

$$\begin{aligned} &\underset{\pi}{\operatorname{argmax}}\, H(\mathbf{A}^T||\mathbf{S}^T) \\ \text{such that: } &\forall_{t,i,a_i^t,a_i^{t\prime},s^{1:t},a_{1:N}^{1:t-1}} \text{ ExpectRegret}_i^\pi(a_i^t, a_i^{t\prime}, s^{1:t}, a_{1:N}^{1:t-1}) \leq 0 \\ &\text{and probabilistic constraints on } \pi. \end{aligned} \tag{A.35}$$

The Lagrangian for the optimization of Equation A.35 when using entire history-dependent probability distributions and parameters is:

$$\Lambda(\pi, \lambda) = H(a_{1:N}^{1:T}||s^{1:T}) - \sum_{t,i,a_i^t,a_i^{t\prime},s^{1:t},a_{1:N}^{1:t-1}} \lambda_{t,i,a_i^t,a_i^{t\prime},s^{1:t},a_{1:N}^{1:t-1}} \text{ ExpectRegret}_i^\pi(a_i^t, a_i^{t\prime}, s^{1:t}, a_{1:N}^{1:t-1}) \tag{A.36}$$

Taking the partial derivative with respect to a history-dependent action probability for a particular state, we have:

$$\begin{aligned} \frac{\partial \Lambda(\pi, \lambda)}{\partial \operatorname{P}(a_{1:N}^t|s^{1:t}, a_{1:N}^{1:t-1})} &= -\sum_{s^{t+1:T}, a_{1:N}^{t+1:T}} \operatorname{P}(s^{t+1:T}, a^{t+1:T}) \log \prod_{\tau=t}^{T} \operatorname{P}(a_{1:N}^\tau|s^{1:\tau}, a_{1:N}^{1:\tau-1}) \\ &\quad - \sum_{i,a_i^{t\prime}} \lambda_{t,i,a_i^t,a_i^{t\prime},s^{1:t},a_{1:N}^{1:t-1}} \text{ ExpectDevGain}_i^\pi(a_i^t, a_i^{t\prime}, s^{1:t}, a_{1:N}^{1:t-1}) \\ &= -\log \operatorname{P}(a_{1:N}^t|s^{1:t}, a_{1:N}^{1:t-1}) - \sum_{s^{t+1:T}, a_{1:N}^{t+1:T}} \operatorname{P}(s^{t+1:T}, a^{t+1:T}) \log \prod_{\tau=t}^{T} \operatorname{P}(a_{1:N}^\tau|s^{1:\tau}, a_{1:N}^{1:\tau-1}) \\ &\quad - \sum_{i,a_i^{t\prime}} \lambda_{t,i,a_i^t,a_i^{t\prime},s^{1:t},a_{1:N}^{1:t-1}} \text{ ExpectDevGain}_i^\pi(a_i^t, a_i^{t\prime}, s^{1:t}, a_{1:N}^{1:t-1}). \end{aligned} \tag{A.37}$$

Equating Equation A.37 to zero provides the form of the history dependent distribution:

$$\begin{aligned} \operatorname{P}(a_{1:N}^t|s^{1:t}, a_{1:N}^{1:t-1}) \propto \exp\Big\{ &\sum_{s^{t+1:T}, a_{1:N}^{t+1:T}} \operatorname{P}(s^{t+1:T}, a^{t+1:T}) \log \prod_{\tau=t}^{T} \operatorname{P}(a_{1:N}^\tau|s^{1:\tau}, a_{1:N}^{1:\tau-1}) \\ &- \sum_{i,a_i^{t\prime}} \lambda_{t,i,a_i^t,a_i^{t\prime},s^{1:t},a_{1:N}^{1:t-1}} \text{ ExpectDevGain}_i^\pi(a_i^t, a_i^{t\prime}, s^{1:t}, a_{1:N}^{1:t-1}) \Big\}. \end{aligned} \tag{A.38}$$

It is important to note that duality in this optimization relies on a feasible solution on the relative interior of the constraint set. This can be accomplished by adding an infinitesimally small amount of slack, ϵ, to the constraint set Ortiz et al. (2007).

Following the argument that the MCECE is Markovian (Lemma 8.14), Equation A.38 reduces to the Markovian form of the theorem. □

A.6 Chapter 9 Proofs

Theorem 9.1. *The expected action visitation frequencies, $D_{a_{x,y}}$, for origin A and goal state B in the deterministic dynamics maximum causal entropy statistic-matching model can be computed from partition functions using the following equation:*

$$D_{a_{x,y}} = \frac{Z_{A\to x} \;\; e^{-cost(a_{x,y})} \;\; Z_{y\to B}}{Z_{A\to B}} \tag{A.39}$$

Proof. Our proof relies on employing the partition function excluding a particular action, which we denote $Z_{A\to B/a_{x\to y}} = \sum_{\zeta_{A\to B}:a_{x,y}\notin\zeta_{A\to B}} e^{-cost(\zeta)}$, the partition function restricted to contain a particular action, which we denote $Z_{A\to B\ni a_{x\to y}} = \sum_{\zeta_{A\to B}:a_{x,y}\in\zeta_{A\to B}} e^{-cost(\zeta)}$, and two mathematical series formulas:

$$\sum_{i=0}^{\infty} x^i = \frac{1}{1-x} \tag{A.40}$$

$$\sum_{i=1}^{\infty} i x^i = \frac{x}{(1-x)^2} \tag{A.41}$$

We can combine these ideas by expressing $Z_{y\to y\ni a_{x,y}}$ as the combination of independently chosen paths to state x and actions $a_{x,y}$. First, we note a useful relationship between action inclusive partition functions and unconstrained partition functions (Equation A.42). We re-express this partition function as a sum of increasingly repeated cycles from y to x and containing action $a_{x,y}$. We finally employ our mathematical series formula (Equation A.40) to obtain Equation A.43.

$$\begin{aligned} Z_{y\to y\ni a_{x,y}} &= e^{-cost(a_{x,y})} Z_{x\to y} = e^{-cost(a_{x,y})} Z_{y\to x} && \text{(A.42)} \\ &= \sum_{k=1}^{\infty} \left(e^{-cost(a_{x,y})} Z_{y\to x/a_{x,y}}\right)^k \\ &= \frac{1}{1 - e^{-cost(a_{x,y})} Z_{y\to x/a_{x,y}}} - 1 && \text{(A.43)} \end{aligned}$$

We can express the regular partition function using the action excluding and action containing partition functions. We can simply combine the partition function for all paths that do not include $a_{x,y}$ with the partition function for all possible paths that do contain $a_{x,y}$ (by taking action $a_{x,y}$ and then returning to S_x). This value can then be expressed in terms of the action inclusion partition function as shown in Equation A.45 by employing Equation A.42

$$\begin{aligned} Z_{A\to x} &= Z_{A\to x/a_{x,y}} + Z_{A\to x/a_{x,y}} \; e^{-cost(a_{x,y})} \; Z_{y\to x} && \text{(A.44)} \\ &= Z_{A\to x/a_{x,y}} \left(1 + Z_{y\to y\ni a_{x,y}}\right) && \text{(A.45)} \end{aligned}$$

We count the unnormalized probability mass of paths weighted by action membership by combining the mass of different disjoint sets of paths. A valid path from A to B containing action $a_{x,y}$ can be characterized as:

- A sub-path from A to x that does not include $a_{x,y}$
- Action $a_{x,y}$
- Any number of "looping" paths from y to x followed by action $a_{x,y}$
- A sub-path from y to B that does not include $a_{x,y}$

Using this construction, the unnormalized probability mass of paths that take action $a_{x\to y}$ exactly $k \geq 1$ times is:

$$\sum_{\zeta_{A\to B}:\#_{a_{x,y}\in\zeta}=k} e^{\theta^\top \mathbf{f}_\zeta} = Z_{A\to x/a_{x,y}} e^{-cost(a_{x,y})} \left(Z_{y\to x/a_{x,y}} e^{-cost(a_{x,y})}\right)^{k-1} Z_{y\to B/a_{x,y}} \tag{A.46}$$

The expected action visitation frequencies can then be calculated by combining these unnormalized probability masses and weighting by the number of actions $a_{x,y}$ contained in each set.

$$\begin{aligned}
D_{a_{x,y}} &= \frac{\sum_{k=1}^{\infty} k\ Z_{A\to x/a_{x,y}}\ e^{-cost(a_{x,y})} \left(Z_{y\to x/a_{x,y}} e^{-cost(a_{x,y})}\right)^{k-1} Z_{y\to B/a_{x,y}}}{Z_{A\to B}} && \text{(A.47)}\\
&= \frac{Z_{A\to x/a_{x,y}}\ e^{-cost(a_{x,y})}\ Z_{y\to B/a_{x,y}} \left(1 + \sum_{k=1}^{\infty} k\ \left(Z_{y\to x/a_{x,y}} e^{-cost(a_{x,y})}\right)^k\right)}{Z_{A\to B}} \\
&= \frac{Z_{A\to x/a_{x,y}}\ e^{-cost(a_{x,y})}\ Z_{y\to B/a_{x,y}} \left(1 + \frac{Z_{y\to x/a_{x,y}} e^{-cost(a_{x,y})}}{(1-Z_{y\to x/a_{x,y}} e^{-cost(a_{x,y})})^2}\right)}{Z_{A\to B}} && \text{(A.48)}\\
&= \frac{Z_{A\to x/a_{x,y}}\ e^{-cost(a_{x,y})}\ Z_{y\to B/a_{x,y}} + Z_{A\to x}\ e^{-cost(a_{x,y})}\ Z_{y\to B}\ \left(1 + Z_{y\to x/a_{x,y}}\ e^{-cost(a_{x,y})}\right)}{Z_{A\to B}} \\
&= \frac{Z_{A\to x}\ e^{-cost(a_{x,y})}\ Z_{y\to B}}{Z_{A\to B}} && \text{(A.49)}
\end{aligned}$$

We reduce Equation A.47 to Equation A.48 using the mathematical series formula (Equation A.41) and Equations A.43 and A.42. We reduce Equation A.48 by expanding the product and applying Equations A.45. Finally, we obtain Equation A.49 using the same logic as employed for Equation A.44. □

Theorem 9.2. *For matrix* $\boldsymbol{A}$*, if* $\lim_{t\to\infty} \boldsymbol{A}^t = \boldsymbol{0}$*, then* $\sum_{t=0}^{\infty} \boldsymbol{A}^t = (\boldsymbol{I} - \boldsymbol{A})^{-1}$*, where* $\boldsymbol{I}$ *is the identity matrix and* $\boldsymbol{A}^0 = \boldsymbol{I}$*.*

Proof sketch. If $\lim_{t\to\infty} \mathbf{A}^t = 0$, then by definition:

$$\sum_{t=0}^{\infty} \mathbf{A}^t = \mathbf{I} + \mathbf{A} \sum_{t=0}^{\infty} \mathbf{A}^t \tag{A.50}$$

$$\sum_{t=0}^{\infty} \mathbf{A}^t (\mathbf{I} - \mathbf{A}) = \mathbf{I} \tag{A.51}$$

$$\sum_{t=0}^{\infty} \mathbf{A}^t = (\mathbf{I} - \mathbf{A})^{-1}, \tag{A.52}$$

by following simple matrix algebra. □

Theorem 9.5. *In a discounted future rewards setting with a bounded instantaneous reward and bounded number of actions, the partition functions of the maximum causal entropy model with finite rewards are guaranteed to converge.*

Proof sketch. With discount rate γ, Equation 6.5 can be re-expressed as:

$$Q_\theta^{\text{soft}}(a_t, s_t) = \gamma E_{\mathrm{P}(s_{t+1}|s_t,a_t)} \left[\operatorname*{softmax}_{a_t} Q_\theta^{\text{soft}}(a_{t+1}, s_{t+1}) | s_t, a_t \right] + \theta^\top \mathbf{f}_{s_t,a_t} \tag{A.53}$$

For non-convergence to occur, the sum of Q_t values at time step t must grow larger than the sum of $Q_{t+\Delta t}$ values at time step $t + \Delta t$ and approaches infinity as $t \to \infty$. However, as these sums approach infinity, the loss from discount of the sum which bounds the previous timestep's partition mass due to the future, $(1-\gamma)^{\Delta t} \sum_{a,s} Q_\theta^{\text{soft}}(a, s)$, must also grow larger than $\sum_{s,a} \theta^\top \mathbf{f}_{s,a}$, preventing growth to infinity (and non-convergence). □

A.7 Chapter 11 Proofs

Theorem 11.1. *The posterior distribution of goals can be obtained from the log partition functions,* $V_{s_x \to s_y}$*, as follows:*

$$\mathrm{P}(G|\boldsymbol{a}_{1:t}, \boldsymbol{s}_{1:t}) \propto \mathrm{P}(G)\ e^{V_{s_t \to G} - V_{s_1 \to G}}. \tag{A.54}$$

Proof.

$$\begin{aligned}
\mathrm{P}(G|\mathbf{a}_{1:t}, \mathbf{s}_{1:t}) &\propto \mathrm{P}(\mathbf{a}_{1:t}, \mathbf{s}_{1:t}|G)\,\mathrm{P}(G) \\
&= \sum_{\mathbf{a}_{t+1:T}, \mathbf{s}_{t+1:T}} \mathrm{P}(\mathbf{a}_{1:T}, \mathbf{s}_{1:T}|G)\,\mathrm{P}(G) \\
&= \frac{\sum_{\mathbf{a}_{t+1:T}, \mathbf{s}_{t+1:T}} e^{\text{reward}(\mathbf{a}_{1:T}, \mathbf{s}_{1:T})}}{\sum_{\mathbf{a}_{1:T}, \mathbf{s}_{1:T}} e^{\text{reward}(\mathbf{a}_{1:T}, \mathbf{s}_{1:T})}}\ \mathrm{P}(G) \\
&\propto \frac{Z_{s_t \to G}}{Z_{s_1 \to G}}\,\mathrm{P}(G) \\
&= \mathrm{P}(G)\ e^{V_{s_t \to G} - V_{s_1 \to G}}
\end{aligned}$$

□

Theorem 11.2. *Action visitation calculations using final reward values as follows:*

$$\begin{aligned}
\phi(G) &= \log \mathrm{P}(G|a_{1:t}, s_{1:t}) \\
&= V_{s_t \to G} - V_{s_1 \to G} + \log \mathrm{P}(G),
\end{aligned}$$

are equivalent to the goal-probability-weighted visitation calculations:

$$D_{s_x|\boldsymbol{s}_{1:t}, \boldsymbol{a}_{1:t}} = \sum_{\boldsymbol{s}_{t+1:T}, \boldsymbol{a}_{t+1:T}} \left(\sum_{\tau=t+1}^{T} I(s_x = s_\tau) \right) \mathrm{P}(\boldsymbol{a}_{t+1:T}, \boldsymbol{s}_{t+1:T}|\boldsymbol{a}_{1:t}, \boldsymbol{s}_{1:t})$$

of Equation 11.4.

Proof sketch. Writing out the visitation frequencies from state s_t to goals with potentials weighted by $\phi(G)$, we have:

$$D_{s_x|\phi(G)} = \frac{Z_{s_t \to s_x} \sum_G Z_{s_x \to G}\ \mathrm{P}(G|\mathbf{a}_{1:t}, \mathbf{s}_{1:t})}{\sum_G Z_{s_t \to G}\ \mathrm{P}(G|\mathbf{a}_{1:t}, \mathbf{s}_{1:t})}, \tag{A.55}$$

which is equivalent to the goal-probability-weighted visitation calculation of the theorem. □

Bibliography

P. Abbeel, A. Coates, M. Quigley, & A. Y. Ng. An application of reinforcement learning to aerobatic helicopter flight. In *NIPS*, pages 1–8, 2007. 14.3, 14.3.1

P. Abbeel & A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proc. ICML*, pages 1–8, 2004. 1.1, 3.2.2, 3.2.2, 6.6, 6.2.1, 14.3, A.3

P. Ardis & C. Brown. Using conditional random fields for decision-theoretic planning. In *Modeling Decisions for Artificial Intelligence*, 2009. 3.1, 3.1.2, 3.1.2, 3.4.1, 3.4.2

B. Argall, S. Chernova, M. Veloso, & B. Browning. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5):469–483, 2009. 3, 6.2

H. Attias. Planning by probabilistic inference. In *Proc. of the 9th Int. Workshop on Artificial Intelligence and Statistics*, 2003. 3.1

R. Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1(1):67–96, 1974. 8, 8.5

J. Bagnell. *Learning decisions: Robustness, uncertainty, and approximation*. PhD thesis, Carnegie Mellon University, 2004. 3.2.6

C. Baker, J. Tenenbaum, & R. Saxe. Bayesian models of human action understanding. *Advances in Neural Information Processing Systems*, 18:99, 2006. 3.2.5

C. Baker, J. Tenenbaum, & R. Saxe. Goal inference as inverse planning. In *Proc. of the cognitive science society*, 2007. 3.2.5, 11.3

L. Bao & S. Intille. Activity recognition from user-annotated acceleration data. *Pervasive Computing*, pages 1–17, 2004. 3.1

R. Bellman. A Markovian decision process. *Journal of Mathematics and Mechanics*, 6:679–684, 1957. 2.2.1, 2.2.1, 4.2.5, 6.1.1, 6.2.2, 6.3.2, 9.1, 9.3.1

M. Ben-Akiva. Structure of passenger travel demand models. 1974. 3.3.3

M. Bennewitz, W. Burgard, & S. Thrun. Learning motion patterns of persons for mobile service robots. In *Proc. ICRA*, pages 3601–3606, 2002. 13.1

S. Boyd, L. El Ghaoui, E. Feron, & V. Balakrishnan. Linear matrix inequalities in system and control theory. *SIAM*, 15, 1994. 3.2.1

S. Boyd & L. Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004. 5.2.1, A.2

S. Boyd, L. Xiao, & A. Mutapcic. Subgradient methods. *Lecture notes of EE392o, Stanford University, Autumn Quarter*, 2003. 10.2.3

S. Brin & L. Page. The anatomy of a large-scale hypertextual Web search engine* 1. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998. 9.1.2

C. Broyden. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76, 1970. 10.2.4

H. Bui, S. Venkatesh, & G. West. Policy recognition in the abstract hidden Markov model. *Journal of Artificial Intelligence Research*, 17(1):451–499, 2002. 3.1.1

U. Chajewska, D. Koller, & D. Ormoneit. Learning an agent's utility function by observing behavior. In *In Proc. of the 18th Intl Conf. on Machine Learning*, pages 35–42, 2001. 3.2.1

G. Cooper. A method for using belief networks as influence diagrams. In *Workshop on Uncertainty in Artificial Intelligence*, pages 55–63, 1988. 3.1

D. Coppersmith & S. Winograd. Matrix multiplication via arithmetic progressions. *Journal of symbolic computation*, 9(3):251–280, 1990. 9.1.2

M. Costa-Gomes, V. Crawford, & B. Broseta. Cognition and behavior in normal-form games: An experimental study. *Econometrica*, 69(5):1193–1235, 2001. 15.2

T. Cover & J. Thomas. *Elements of information theory*. John Wiley and sons, 2006. 4.1, 4.1.3, 4.16, 4.17, 8.2.2

M. da Silva, F. Durand, & J. Popović. Linear Bellman combination for control of character animation. *ACM Transactions on Graphics (TOG)*, 28(3):1–10, 2009. 3.2.6

E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1: 269–271, 1959. 13.2.1

A. Drake. *Observation of a Markov process through a noisy channel*. PhD thesis, Massachusetts Institute of Technology, 1962. 2.2.1

M. Dudık & G. Gordon. A sampling-based approach to computing equilibria in succinct extensive-form games. In *Proc. 25th Conference on Uncertainty in Artificial Intelligence*. Citeseer, 2009. 8.3

M. Dudík & R. E. Schapire. Maximum entropy distribution estimation with generalized regularization. In *Proc. COLT*, pages 123–138, 2006. 5.1.4, 5.1

E. Feinberg & A. Shwartz. *Handbook of Markov Decision Process Methods and Applications*. Kluwer Academic Publishers, 2002. 2.8, 2.2.1

D. Ferguson & A. T. Stentz. Field D*: An interpolation-based path planner and replanner. In *Proc. ISRR*, pages 239–253, 2005. 13.2.1

J. Filar, K. Vrieze, & O. Vrieze. *Competitive Markov decision processes*. Springer Verlag, 1997. 8.2, 8.2

R. Fletcher. A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317, 1970. 10.2.4

A. F. Foka & P. E. Trahanias. Predictive autonomous robot navigation. In *Proc. IROS*, pages 490–495, 2002. 13.1

N. Gaarder & J. Wolf. The capacity region of a multiple-access discrete memoryless channel can increase with feedback (corresp.). *Information Theory, IEEE Transactions on*, 21(1):100–102, 1975. 4.2.5

A. Galata, N. Johnson, & D. Hogg. Learning variable length Markov models of behaviour. *Computer Vision and Image Understanding (CVIU) Journal*, 2001. 3.1.1, 13.1, 13.3.5

A. Giffin & A. Caticha. Updating probabilities with data and moments. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume 954, pages 74–84, 2007. 5.1.4

I. Gilboa & E. Zemel. Nash and correlated equilibria: Some complexity considerations. *Games and Economic Behavior*, 1(1):80–93, 1989. 8.1.2

D. Goldfarb. A family of variable metric updates derived by variational means. *Mathematics of Computing*, 24(109):23–26, 1970. 10.2.4

Google 2008. Google maps, April 2008. URL `http://maps.google.com/`. 12.2

A. Greenwald & Hall. Correlated Q-learning. In *Proc. ICML*, pages 242–249, 2003. 8.6, 8.1.2

P. D. Grünwald & A. P. Dawid. Game theory, maximum entropy, minimum discrepancy, and robust bayesian decision theory. *Annals of Statistics*, 32:1367–1433, 2003. 5.2, 5.2.3, 8.11, A.2, A.5

P. Hart, N. Nilsson, & B. Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4:100–107, 1968. 9.1.4

S. Hart & A. Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000. 8.2

D. Heckerman, J. S. Breese, & K. Rommelse. Troubleshooting under uncertainty. In *Communications of the ACM*, pages 121–130, 1994. 14.2.2

P. Henry, C. Vollmer, B. Ferris, & D. Fox. Learning to Navigate Through Crowded Environments. In *Proc. ICRA*, 2010. 13.4

R. Howard. Information value theory. *IEEE Transactions on systems science and cybernetics*, 2 (1):22–26, 1966. 6.4

R. A. Howard & J. E. Matheson. Influence diagrams. In *Readings on the Principles and Applications of Decision Analysis*, pages 721–762. Strategic Decisions Group, 1984. 2.2.3, 2.2.3, 7

B. Hull, V. Bychkovsky, Y. Zhang, K. Chen, M. Goraczko, A. K. Miu, E. Shih, H. Balakrishnan, & S. Madden. CarTel: A Distributed Mobile Sensor Computing System. In *ACM SenSys*, pages 125–138, 2006. 12.6.2

E. Jaynes. The relation of Bayesian and maximum entropy methods. *Maximum-entropy and Bayesian methods in science and engineering*, 1, 1988. 5.1.4

E. Jaynes & G. Bretthorst. *Probability theory: the logic of science*. Cambridge Univ Pr, 2003. 5.1.1

E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106:620–630, 1957. 1.1, 1.5, 5, 5.1, 5.1.1, 8.7

F. Jensen, F. Jensen, S. Dittmer, et al. From influence diagrams to junction trees. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, 1994. 3.1

R. Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960. 1

R. Kalman. When is a linear control system optimal? *Trans. ASME, J. Basic Engrg.*, 86:51–60, 1964. 3.2.1

R. E. Kalman & R. S. Bucy. New results in linear filtering and prediction theory. *AMSE Journ. Basic Eng.*, pages 95–108, 1962. 13.1

B. Kappen, V. Gomez, M. Opper, & T. Berlin. Optimal control as a graphical model inference problem. *(In submission)*, 2010. 3.1

H. Kappen. Linear theory for control of nonlinear stochastic systems. *Physical review letters*, 95 (20):200201, 2005. 3.2.6

R. Kindermann, J. Snell, & A. M. Society. *Markov random fields and their applications*. American Mathematical Society Providence, Rhode Island, 1980. 2.1.2

J. Kivinen & M. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997. 10.2.2

G. Kramer. *Directed Information for Channels with Feedback*. PhD thesis, Swiss Federal Institute of Technology (ETH) Zurich, 1998. 4.2.2, 4.18, 4.19, 4.2.5, 4.26, 4.2.5, 5.3.1, 8.2.1

A. Krause & C. Guestrin. Optimal nonmyopic value of information in graphical models-efficient algorithms and theoretical limits. In *International Joint Conference on Artificial Intelligence*, volume 19, page 1339. Citeseer, 2005. 6.4

D. Krishnamurthy & E. Todorov. Inverse Optimal Control with Linearly-solvable MDPs. In *ICML*, 2010. 3.2.6, 3.4.2

J. Krumm. A Markov model for driver route prediction. *Society of Automative Engineers (SAE) World Congress*, 2008. 3, 12.6.1

J. Krumm & E. Horvitz. Predestination: Inferring destinations from partial trajectories. In *Proc. Ubicomp*, pages 243–260, 2006. 12.6.3

S. Kullback & R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:49–86, 1951. 4.1.1

S. Kumar & M. Hebert. Discriminative random fields. *Int. J. Comput. Vision*, 68(2):179–201, 2006. ISSN 0920-5691. 2.1.2, 3.1.2

J. Lafferty, A. McCallum, & F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*, pages 282–289, 2001. 1.1, 2.1.2, 2.1.2, 3.1.2, 5.3.4, 6, 6.2.5, 6.2.5

C. Lemke & J. Howson Jr. Equilibrium points of bimatrix games. *Journal of the Society for Industrial and Applied Mathematics*, 12(2):413–423, 1964. 8.1.2

J. Letchner, J. Krumm, & E. Horvitz. Trip router with individualized preferences (trip): Incorporating personalization into route planning. In *Proc. IAAI*, pages 1795–1800, 2006. 12.2, 12.6.2

L. Liao, D. Fox, & H. Kautz. Extracting places and avctivities from GPS traces using hierarchical conditional random fields. *Int. J. Rob. Res.*, 26(1):119–134, 2007a. ISSN 0278-3649. 2.1.2, 3.1.2

L. Liao, D. J. Patterson, D. Fox, & H. Kautz. Learning and inferring transportation routines. *Artificial Intelligence*, 171(5-6):311–331, 2007b. ISSN 0004-3702. 12.6.1

R. Luce. *Individual choice behavior*. Wiley New York, 1959. 3.3.1

L. MacDermed & C. L. Isbell. Solving Stochastic Games. In *Proc. NIPS*, 2009. 8.3.2

R. MacLachlan. Tracking moving objects from a moving vehicle using a laser scanner. Technical Report CMU-RI-TR-05-07, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, 2005. 13.3.1

R. Madhavan & C. Schlenoff. Moving object prediction for off-road autonomous navigation. In *SPIE Aerosense Conference*, 2003. 13.1

MapPoint 2008. Microsoft mappoint, April 2008. URL `http://www.microsoft.com/mappoint/`. 12.2

MapQuest 2008. Mapquest, April 2008. URL `http://www.mapquest.com/`. 12.2

H. Marko. The bidirectional communication theory – a generalization of information theory. In *IEEE Transactions on Communications*, pages 1345–1351, 1973. 4.2.2

J. L. Massey. Causality, feedback and directed information. In *Proc. IEEE International Symposium on Information Theory and Its Applications*, pages 27–30, 1990. 4.2.2, 4.23, 4.24, 4.2.5, 4.25, 4.2.5

D. McFadden. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, pages 105–142, 1974. 3.3.1, 3.3.2, 3.4.1, 3.4.2

R. McKelvey & T. Palfrey. An experimental study of the centipede game. *Econometrica: Journal of the Econometric Society*, 60(4):803–836, 1992. 15.2

R. McKelvey & T. Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995. 15.1

R. McKelvey & T. Palfrey. Quantal response equilibria for extensive form games. *Experimental Economics*, 1(1):9–41, 1998. 15.1

C. Mertz. A 2d collision warning framework based on a Monte Carlo approach. In *Proc. of Intelligent Transportation Systems*, 2004. 13.1

A. C. Miller, M. M. Merkhofer, R. A. Howard, J. E. Matheson, & T. R. Rice. Development of automated aids for decision analysis. Technical report, 1976. 2.2.3, 7

K. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, UNIVERSITY OF CALIFORNIA, 2002. 2.2

C. Murray & G. Gordon. Multi-robot negotiation: approximating the set of subgame perfect equilibria in general-sum stochastic games. In *Proc. NIPS*, pages 1001–1008, 2007. 8.3.2

J. Nash. Non-cooperative games. *Annals of mathematics*, 54(2):286–295, 1951. 8, 8.1.2, 8.1.2, 8.1.2

Y. Nesterov & A. Nemirovski. *Interior-point polynomial algorithms in convex programming*. Society for Industrial Mathematics, 1987. 10.2.4

G. Neu & C. Szepesvári. Apprenticeship learning using inverse reinforcement learning and gradient methods. In *Proc. UAI*, pages 295–302, 2007. 3.2.5, 3.6, 12.6.2

A. Y. Ng & S. Russell. Algorithms for inverse reinforcement learning. In *Proc. ICML*, pages 663–670, 2000. 1.1, 3.2.1, 3.1, 3.2.1

T. Nielsen & F. Jensen. Learning a decision maker's utility function from (possibly) inconsistent behavior. *Artificial Intelligence*, 160(1-2):53–78, 2004. 3.2.7, 7.4

A. Nilim & L. El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005. 3.2.6

L. E. Ortiz, R. E. Shapire, & S. M. Kakade. Maximum entropy correlated equilibrium. In *AISTATS*, pages 347–354, 2007. 1.1, 8.7, 8.1.2, 8.2.1, 8.10, A.5

C. Papadimitriou & T. Roughgarden. Computing equilibria in multi-player games. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 82–91. Society for Industrial and Applied Mathematics, 2005. 8.1.2

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000. 4.2.1

H. H. Permuter, Y.-H. Kim, & T. Weissman. On directed information and gambling. In *Proc. IEEE International Symposium on Information Theory*, pages 1403–1407, 2008. 4.19, 4.2.5, 4.2.5, 4.27

S. Petti & T. Fraichard. Safe motion planning in dynamic environments. In *Proc. IROS*, pages 3726–3731, 2005. 13.1

K. Poh & E. Horvitz. A graph-theoretic analysis of information value. In *Uncertainty in Artificial Intelligence: Proceedings of the Twelfth Conference*, pages 427–435, 1996. 6.4

D. Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In *Advances in Neural Information Processing Systems 1*, 1989. 3.2

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, 1994. 1, 2.2.1, 12.3.1

D. Ramachandran & E. Amir. Bayesian inverse reinforcement learning. In *Proc. IJCAI*, pages 2586–2591, 2007. 3.2.5, 12.6.2

N. Ratliff, J. A. Bagnell, & M. Zinkevich. Maximum margin planning. In *Proc. ICML*, pages 729–736, 2006. 3.2.3, 12.6.2, 14.3, 14.3.2

N. Ratliff, D. Bradley, J. Bagnell, & J. Chestnutt. Boosting structured prediction for imitation learning. *Advances in Neural Information Processing Systems*, 19:1153, 2007. 3.2.3

N. Ratliff, B. Ziebart, K. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey, & S. Srinivasa. Inverse optimal heuristic control for imitation learning. In *Proc. AISTATS*, pages 424–431, 2009. 1.7

H. Robbins & S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951. 10.2.2

A. Rosenblueth, N. Wiener, & J. Bigelow. Behavior, purpose and teleology. *Philosophy of Science*, 10(1):18–24, 1943. 1.2, 1.3, 1.3, 1.3, 1.3

C. Schlenoff, R. Madhavan, & T. Barbera. A hierarchical, multi-resolutional moving object prediction approach for autonomous on-road driving. *Proc. ICRA*, 2:1956–1961, 2004. ISSN 1050-4729. 13.1

R. Shachter. Evaluating influence diagrams. *Operations Research*, 34(6):871–882, 1986. 2.2.3

R. Shachter. Model building with belief networks and influence diagrams. 2007. 2.2

R. Shachter & M. Peot. Decision making using probabilistic inference methods. In *Uncertainty in Artificial Intelligence*, pages 276–283, 1992. 2.2.3, 3.1

D. Shanno. Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24(111):647–656, 1970. 10.2.4

C. Shannon. The zero error capacity of a noisy channel. *Information Theory, IRE Transactions on*, 2(3):8–19, 1956. 4.2.5

C. Shannon. Two-way communication channels. In *Proc. 4th Berkeley Symp. Math. Stat. Prob*, volume 1, pages 351–384, 1961. 4.2.5

C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27, 1948. 4

N. Shor, K. Kiwiel, & A. Ruszcayski. *Minimization methods for non-differentiable functions*. Springer-Verlag New York, Inc. New York, NY, USA, 1985. 10.2.3

R. Simmons, B. Browning, Y. Zhang, & V. Sadekar. Learning to predict driver route and destination intent. *Proc. Intelligent Transportation Systems Conference*, pages 127–132, 2006. 3, 12.6.1, 12.6.2

D. Stahl & P. Wilson. On Players Models of Other Players: Theory and Experimental Evidence. *Games and Economic Behavior*, 10(1):218–254, 1995. 15.2

A. Steinfeld, D. Manes, P. Green, & D. Hunter. Destination entry and retrieval with the ali-scout navigation system. *The University of Michigan Transportation Research Institute No. UMTRI-96-30*, 1996. 12.1, 12.6.3

R. J. Sternberg & W. Salter. *Handbook of Human Intelligence*. Cambridge University Press, 1982. 1, 1.3

U. Syed, M. Bowling, & R. E. Schapire. Apprenticeship learning using linear programming. In *ICML '08: Proceedings of the 25th international conference on Machine learning*, pages 1032–1039, 2008. 3.2.2, 3.2.4

U. Syed & R. Schapire. A game-theoretic approach to apprenticeship learning. In *Advances in Neural Information Processing Systems 20*, pages 1449–1456. 2007. 3.2.4, 3.5

S. Tadokoro, M. Hayashi, Y. Manabe, Y. Nakami, & T. Takamori. On motion planning of mobile robots which coexist and cooperate with human. In *Proc. IROS*, pages 518–523, 1995. 13, 13.1

E. M. Tapia, S. S. Intille, & K. Larson. Activity recognition in the home using simple and ubiquitous sensors. In *Proc. Pervasive*, pages 158–175, 2004. 3.1.1

S. Tatikonda. *Control under Communication Constraints*. PhD thesis, Massachusetts Institute of Technology, 2000. 4.2.5, 4.2.5

R. Taylor. Purposeful and non-purposeful behavior: A rejoinder. *Philosophy of Science*, 17(4): 327–332, 1950. 1.3, 1.1

S. Thrun, W. Burgard, & D. Fox. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, September 2005. 12.4.3, 13.1

N. Tishby & D. Polani. Information Theory of Decisions and Actions. 2010. 6.4

E. Todorov. Linearly-solvable Markov decision problems. In *NIPS*, pages 1369–1376, 2006. 3.2.6, 3.4.1, 3.4.2

E. Todorov. Compositionality of optimal control laws. In *Advances in Neural Information Processing Systems*, 2009a. 3.2.6, 3.8, 11.3

E. Todorov. Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences*, 106(28):11478, 2009b. 3.2.6

M. Toussaint. Robot trajectory optimization using approximate inference. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM New York, NY, USA, 2009. 3.1.2, 3.4.1, 3.4.2, 6.17

M. Toussaint & A. Storkey. Probabilistic inference for solving discrete and continuous state Markov Decision Processes. In *Proceedings of the 23rd international conference on Machine learning*, page 952. ACM, 2006. 3.1, 3.1.1

K. Tsuda, G. Ratsch, & M. Warmuth. Matrix exponentiated gradient updates for on-line learning and bregman projection. *Journal of Machine Learning Research*, 6(1):995, 2006. 10.3.1

D. L. Vail, M. M. Veloso, & J. D. Lafferty. Conditional random fields for activity recognition. In *Proc. AAMAS*, pages 1–8, 2007. 2.1.2, 3.1.2

D. A. Vasquez Govea, T. Fraichard, O. Aycard, & C. Laugier. Intentional motion on-line learning and prediction. In *Proc. Int. Conf. on Field and Service Robotics*, pages 411–425, 2005. 13.1

M. Veloso. *Planning and learning by analogical reasoning*. Springer, 1994. 3.1.1

D. Verma & R. Rao. Goal-based imitation as probabilistic inference over graphical models. In *Proc. NIPS*, pages 1393–1400, 2006. 3.1.1

S. Vishwanathan, N. Schraudolph, M. Schmidt, & K. Murphy. Accelerated training of conditional random fields with stochastic gradient methods. In *Proceedings of the 23rd international conference on Machine learning*, page 976. ACM, 2006. 10.2.2

M. Weiser. The computer for the twenty-first century. *Scientific American*, 265(3):94–104, 1991. 1.2

J. R. Wright & K. Leyton-Brown. Beyond equilibrium: Predicting human behavior in normal-form games. In *Proc. AAAI*, 2010. 15.2

Yahoo 2008. Yahoo! maps, April 2008. URL `http://maps.yahoo.com/`. 12.2

N. Zhang. Probabilistic inference in influence diagrams. *Computational Intelligence*, 14(4):475–497, 1998. 3.1

B. D. Ziebart, J. A. Bagnell, & A. K. Dey. Probabilistic approaches for robotics and control. In *NIPS Workshop: Probabilistic Approaches for Robotics and Control*, 2009a. 1.7

B. D. Ziebart, J. A. Bagnell, & A. K. Dey. Maximum Causal Entropy Correlated Equilibria for Markov Games. In *AAAI Worksop: Interactive Decision Theory and Game Theory*, 2010a. 1.1, 1.7

B. D. Ziebart, J. A. Bagnell, & A. K. Dey. Modeling interaction via the principle of maximum causal entropy. In *ICML*, 2010b. 1.1, 1.7, 5, 6

B. D. Ziebart, A. K. Dey, & J. A. Bagnell. Fast planning for dynamic preferences. In *Proc. ICAPS*, 2008a. 1.7

B. D. Ziebart, A. Maas, J. A. Bagnell, & A. K. Dey. Maximum entropy inverse reinforcement learning. In *NIPS Workshop: Robotics Challenges for Machine Learning*, 2007. 1.7

B. D. Ziebart, A. Maas, J. A. Bagnell, & A. K. Dey. Maximum entropy inverse reinforcement learning. In *Proc. AAAI*, 2008b. 1.1, 1.7, 3.1.2, 3.2.6, 3.4.2

B. D. Ziebart, A. Maas, A. K. Dey, & J. A. Bagnell. Navigate like a cabbie: Probabilistic reasoning from observed context-aware behavior. In *Proc. Ubicomp*, 2008c. 1.7, 3.2.6

B. D. Ziebart, N. Ratliff, G. Gallagher, C. Mertz, K. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey, & S. Srinivasa. Planning-based prediction for pedestrians. In *Proc. IROS*, 2009b. 1.7

M. Zinkevich, A. Greenwald, & M. Littman. Cyclic equilibria in Markov games. In *Proc. NIPS*, volume 18, pages 1641–1648, 2006. 8.3.2, 14.1.1